Wine Analysis in Vinho Verde Wine

MS-AAI-500
Team 4 Final Project

Paul Parks, Alden Caterio, Mayank Bhatt

Introduction

Introduction

Wine quality is a highly subjective but heavily studied topic.

Many efforts are dedicated to finding out which factors contribute to high quality wine.

In our report, we will use **statistical analysis** to analyze wine from a popular Portuguese wine company, ***Vinho Verde***.

VINHO
VERDE

Objective

Our objective is to determine how **wine physicochemical attributes** affect **quality**.

We will provide a **statistically-validated strategy to improve wine quality ratings**.

Our report will include:

- Data to support the correlation between attributes and higher wine quality.
- Actionable recommendations on how to improve wine quality.

Target Audience

Our target audience is **business leaders in the Wine industry** who wish to improve the overall quality of their wine products.

Understanding **how to increase wine quality** can have **positive benefits** for wine companies and their constituents such as:

1. **Increased profits from wine sales**
2. **Improved production processes**

Data Cleaning/Preparation

Exploratory Data Analysis

Data Cleaning and Preparation

General Descriptive Statistics

- Two Datasets:
  a. Red wine
  b. White wine
- The red wine dataset contains 1599 samples.
- The white wine dataset contains 4898 samples.
- Each dataset contains 12 characteristics:
  a. 11 physicochemical attributes (explanatory variables)
  b. 1 "Quality" attribute (dependent variable)

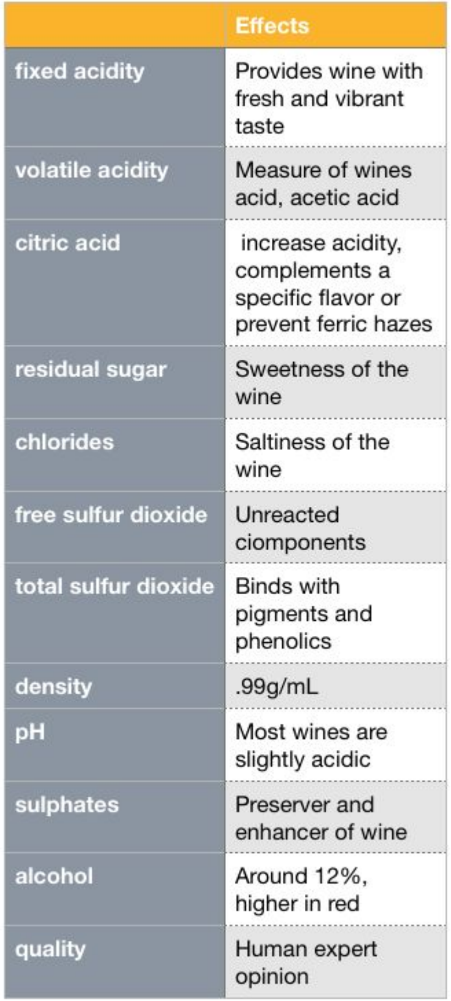
Data Cleaning and Preparation

- 2 types of wine: red and white
- Physicochemical attributes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
- Data provided by UCI Machine Learning Repository
- Data gathered by physicochemical laboratory test
- Sensory test to determine quality

Physicochemical Attributes Explained

Sweetness
Acidity
Tannin
Alcohol
Body
wine characteristics
WINE FOLLY

Table 1



	Effects
fixed acidity	Provides wine with fresh and vibrant taste
volatile acidity	Measure of wines acid, acetic acid
citric acid	increase acidity, complements a specific flavor or prevent ferric hazes
residual sugar	Sweetness of the wine
chlorides	Saltiness of the wine
free sulfur dioxide	Unreacted ciomponents
total sulfur dioxide	Binds with pigments and phenolics
density	.99g/mL
pH	Most wines are slightly acidic
sulphates	Preserver and enhancer of wine
alcohol	Around 12%, higher in red
quality	Human expert opinion

White Wine Attributes

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
Count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267	5.877909
std	0.843868	0.100795	0.12102	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621	0.885639
min	3.8	0.08	0	0.6	0.009	2	9	0.98711	2.72	0.22	8	3
25%	6.3	0.21	0.27	1.7	0.036	23	108	0.991723	3.09	0.41	9.5	5
50%	6.8	0.26	0.32	5.2	0.043	34	134	0.99374	3.18	0.47	10.4	6
75%	7.3	0.32	0.39	9.9	0.05	46	167	0.9961	3.28	0.55	11.4	6
max	14.2	1.1	1.66	65.8	0.346	289	440	1.03898	3.82	1.08	14.2	9

Red Wine Attributes

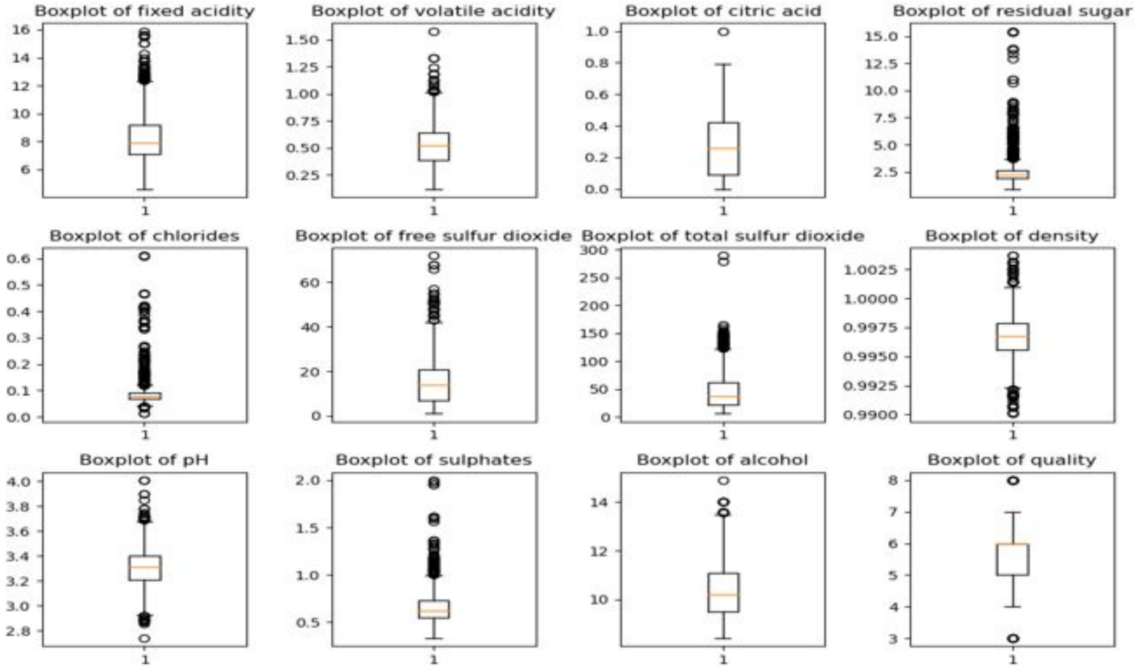
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Exploratory Data Analysis

Boxplots help us visualize the following statistics:

- Average
- Variation
- Minimum
- Maximum
- Outliers

Outliers (dark circles) are present in many of the attributes.



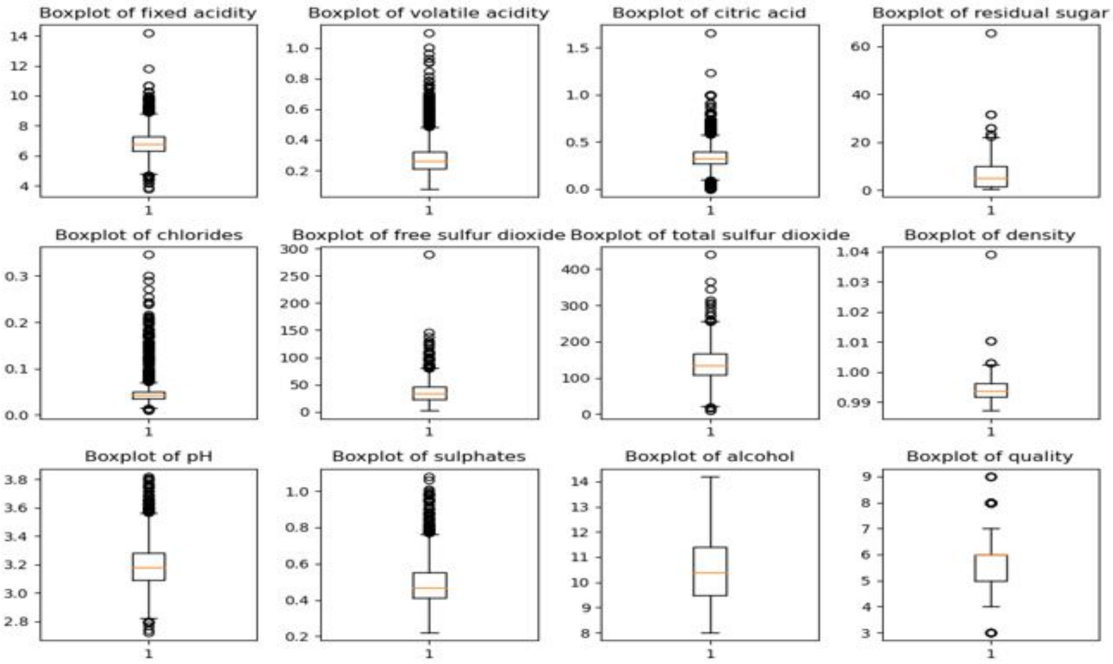
Boxplots for Red Wine Attributes

Exploratory Data Analysis

Boxplots help us visualize the following statistics:

- Average
- Variation
- Minimum
- Maximum
- Outliers

Outliers (dark circles) are present in many of the attributes.



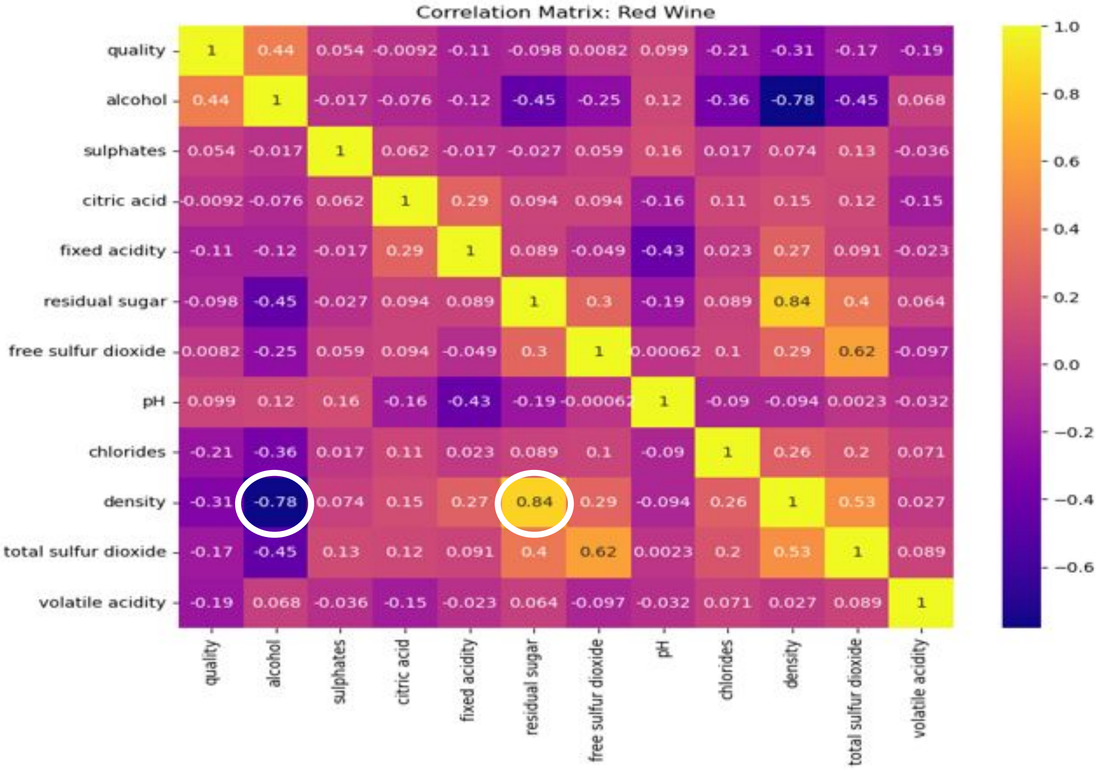
Boxplots for White Wine Attributes

Exploratory Data Analysis

High correlation between attributes can also impact the results of statistical models.

A **correlation matrix** helps visualize which variables are correlated.

- Density and alcohol are negatively correlated
- Density and residual sugar are positively correlated



Correlation Matrix: Red Wine

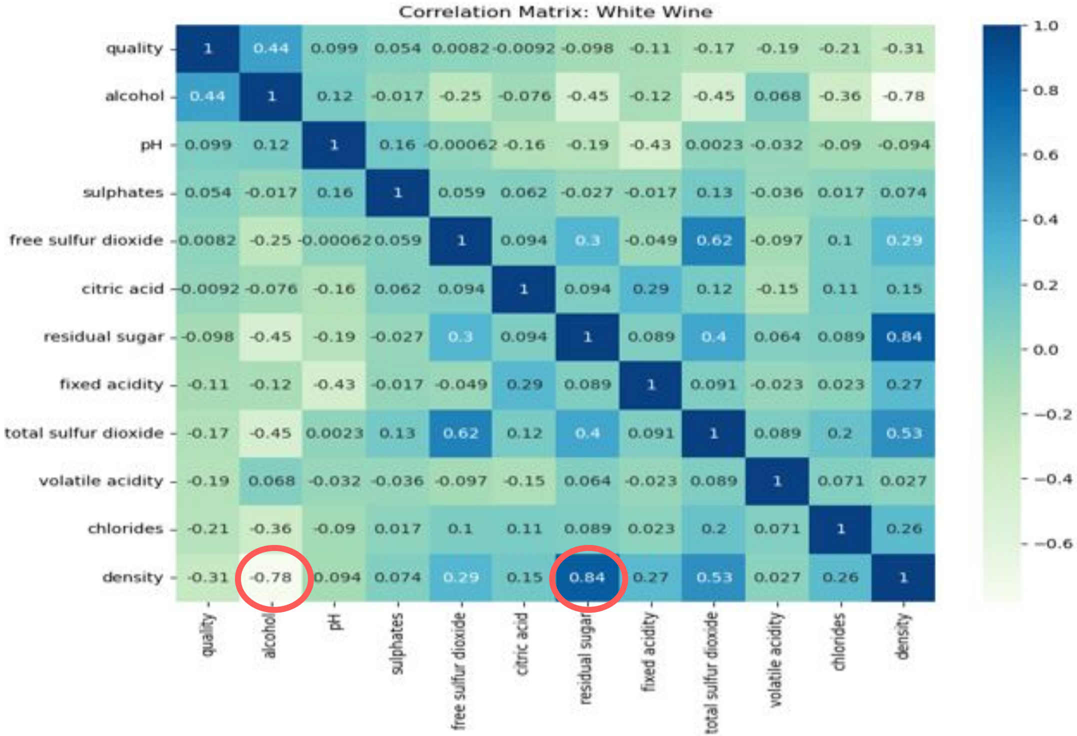
	quality	alcohol	sulphates	citric acid	fixed acidity	residual sugar	free sulfur dioxide	pH	chlorides	density	total sulfur dioxide	volatile acidity
quality	1	0.44	0.054	-0.0092	-0.11	-0.098	0.0082	0.099	-0.21	-0.31	-0.17	-0.19
alcohol	0.44	1	-0.017	-0.076	-0.12	-0.45	-0.25	0.12	-0.36	-0.78	-0.45	0.068
sulphates	0.054	-0.017	1	0.062	-0.017	-0.027	0.059	0.16	0.017	0.074	0.13	-0.036
citric acid	-0.0092	-0.076	0.062	1	0.29	0.094	0.094	-0.16	0.11	0.15	0.12	-0.15
fixed acidity	-0.11	-0.12	-0.017	0.29	1	0.089	-0.049	-0.43	0.023	0.27	0.091	-0.023
residual sugar	-0.098	-0.45	-0.027	0.094	0.089	1	0.3	-0.19	0.089	0.84	0.4	0.064
free sulfur dioxide	0.0082	-0.25	0.059	0.094	-0.049	0.3	1	0.00062	0.1	0.29	0.62	-0.097
pH	0.099	0.12	0.16	-0.16	-0.43	-0.19	-0.00062	1	-0.09	-0.094	0.0023	-0.032
chlorides	-0.21	-0.36	0.017	0.11	0.023	0.089	0.1	-0.09	1	0.26	0.2	0.071
density	-0.31	-0.78	0.074	0.15	0.27	0.84	0.29	-0.094	0.26	1	0.53	0.027
total sulfur dioxide	-0.17	-0.45	0.13	0.12	0.091	0.4	0.62	0.0023	0.2	0.53	1	0.089
volatile acidity	-0.19	0.068	-0.036	-0.15	-0.023	0.064	-0.097	-0.032	0.071	0.027	0.089	1

Exploratory Data Analysis

High correlation between attributes can also impact the results of statistical models.

A **correlation matrix** helps visualize which variables are correlated.

- Density and alcohol are negatively correlated
- Density and residual sugar are positively correlated



Model Selection

Model Selection

The statsmodels GLM (Generalized Linear Model) is a statistical tool that helps us understand and predict the relationship between different factors or variables.

- Our model allows us to predict a Wine quality score given the Wine’s physicochemical test data.

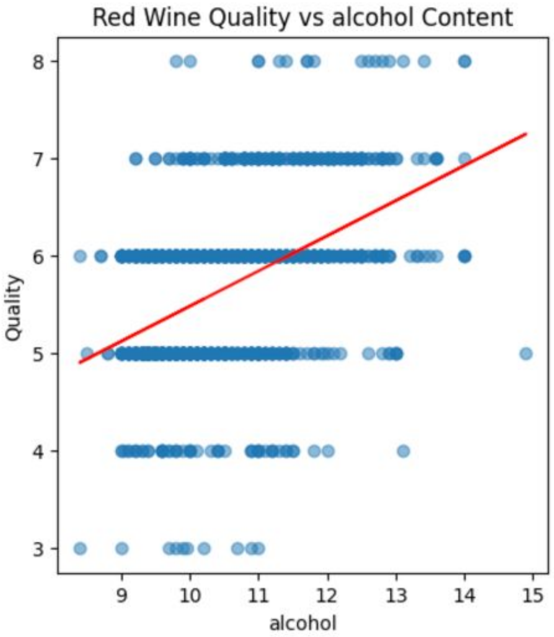
Model Analysis

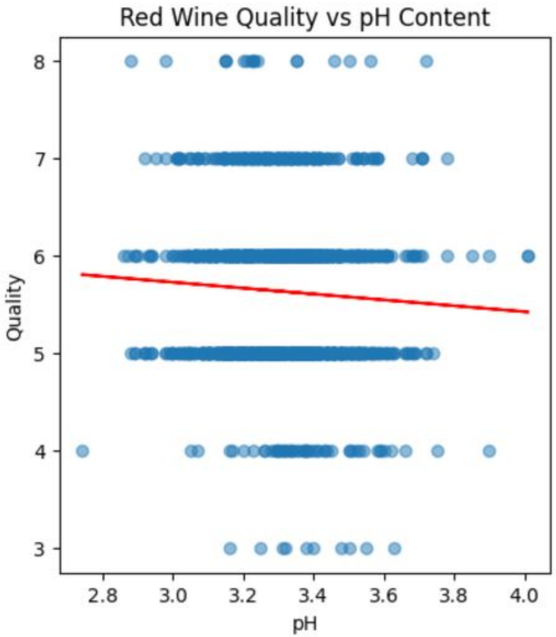
Model Analysis - Red Wine

- Trained on **1,124** observations

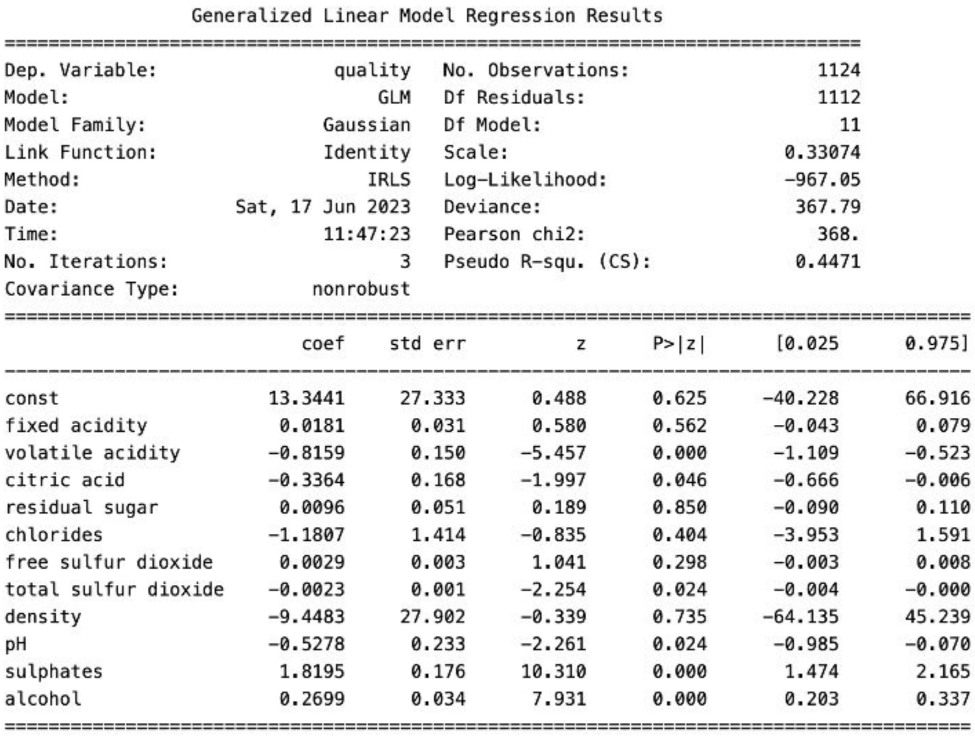
- **'sulphates'** and **'alcohol'** appear to have a significant positive effect on wine quality

- **'volatile acidity'**, **'citric acid'**, **'total sulfur dioxide'**, and **'pH'** also significantly influence but negatively


Red Wine Quality vs alcohol Content


Red Wine Quality vs pH Content

Model Analysis - Red Wine

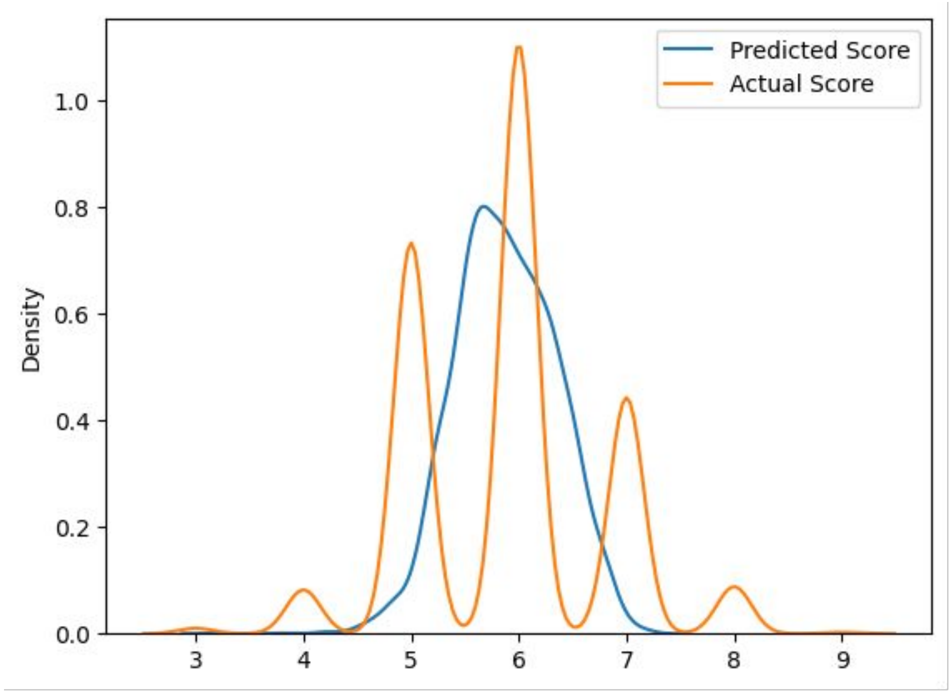


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                quality   No. Observations:                 1124
Model:                            GLM   Df Residuals:                     1112
Model Family:                Gaussian   Df Model:                           11
Link Function:               Identity   Scale:                         0.33074
Method:                          IRLS   Log-Likelihood:                -967.05
Date:                Sat, 17 Jun 2023   Deviance:                       367.79
Time:                        11:47:23   Pearson chi2:                     368.
No. Iterations:                      3   Pseudo R-squ. (CS):             0.4471
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   13.3441     27.333      0.488      0.625     -40.228      66.916
fixed acidity            0.0181      0.031      0.580      0.562      -0.043       0.079
volatile acidity        -0.8159      0.150     -5.457      0.000      -1.109      -0.523
citric acid             -0.3364      0.168     -1.997      0.046      -0.666      -0.006
residual sugar           0.0096      0.051      0.189      0.850      -0.090       0.110
chlorides               -1.1807      1.414     -0.835      0.404      -3.953       1.591
free sulfur dioxide      0.0029      0.003      1.041      0.298      -0.003       0.008
total sulfur dioxide    -0.0023      0.001     -2.254      0.024      -0.004      -0.000
density                 -9.4483     27.902     -0.339      0.735     -64.135      45.239
pH                      -0.5278      0.233     -2.261      0.024      -0.985      -0.070
sulphates                1.8195      0.176     10.310      0.000       1.474       2.165
alcohol                  0.2699      0.034      7.931      0.000       0.203       0.337
========================================================================================
```

- Our model can predict 44% of the quality from the independent variables

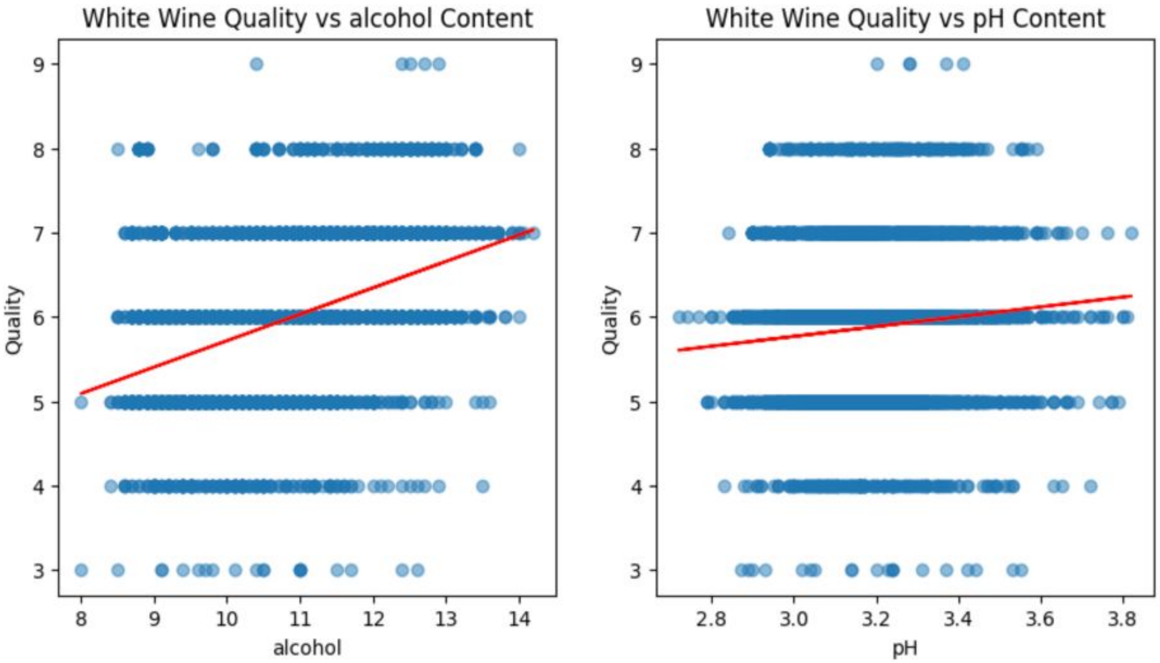
- volatile acidity, citric acid, total sulpher dioxide, pH, sulphates, and alcohol are all statistically significant

Red Wine Model with Predictions

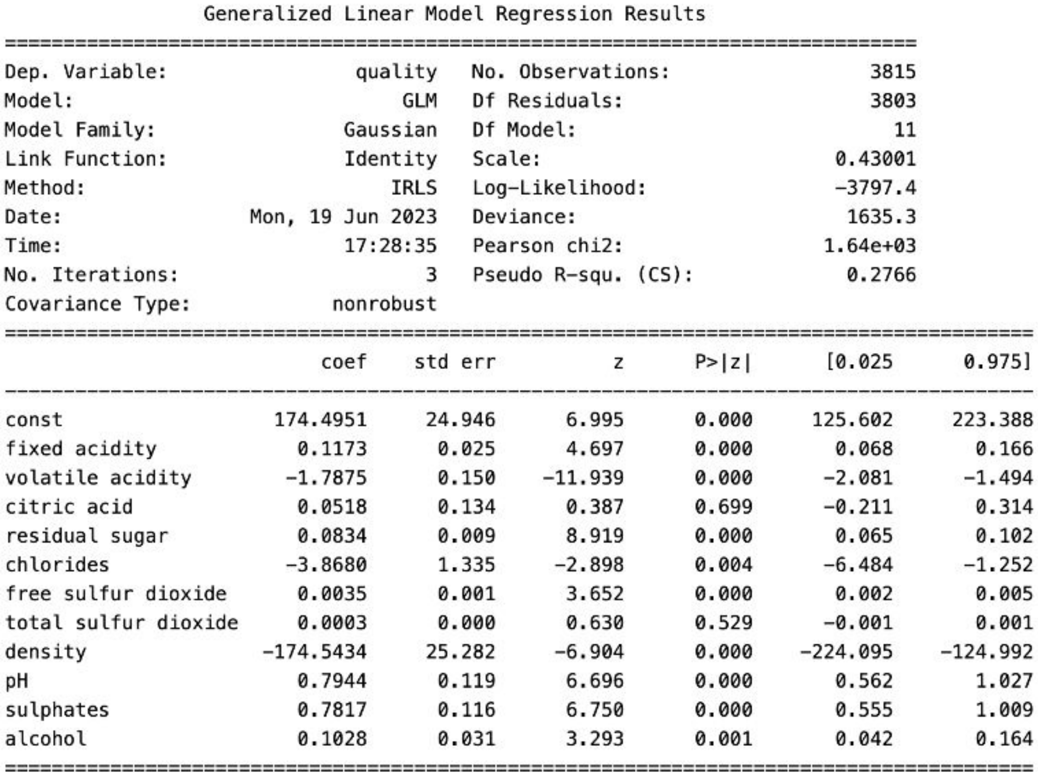


Model Analysis - White Wine

- Trained on **3,815** observations
- **'sulphates'** and **'alcohol'** appear to have a significant positive effect on wine quality
- **'residual sugar'** has a positive relationship with wine quality
- **'density'** and **'volatile acidity'** exhibit a negative relationship with quality



Model Analysis - White Wine

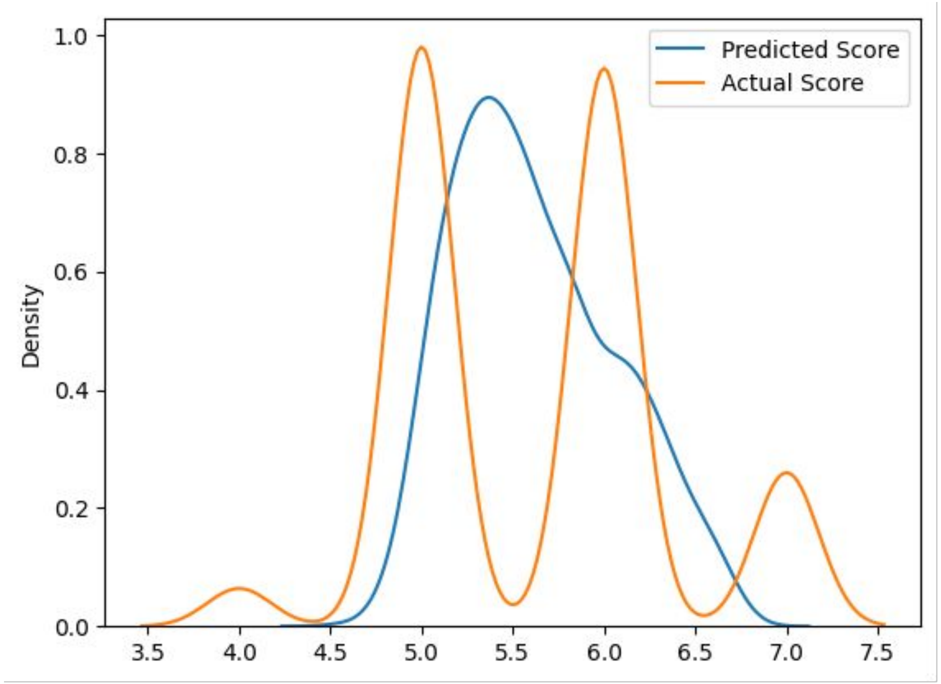


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                quality   No. Observations:                 3815
Model:                            GLM   Df Residuals:                     3803
Model Family:                Gaussian   Df Model:                           11
Link Function:               Identity   Scale:                         0.43001
Method:                          IRLS   Log-Likelihood:                -3797.4
Date:                Mon, 19 Jun 2023   Deviance:                       1635.3
Time:                        17:28:35   Pearson chi2:                 1.64e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.2766
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  174.4951     24.946      6.995      0.000     125.602     223.388
fixed acidity            0.1173      0.025      4.697      0.000       0.068       0.166
volatile acidity        -1.7875      0.150    -11.939      0.000      -2.081      -1.494
citric acid              0.0518      0.134      0.387      0.699      -0.211       0.314
residual sugar           0.0834      0.009      8.919      0.000       0.065       0.102
chlorides               -3.8680      1.335     -2.898      0.004      -6.484      -1.252
free sulfur dioxide      0.0035      0.001      3.652      0.000       0.002       0.005
total sulfur dioxide     0.0003      0.000      0.630      0.529      -0.001       0.001
density               -174.5434     25.282     -6.904      0.000    -224.095    -124.992
pH                       0.7944      0.119      6.696      0.000       0.562       1.027
sulphates                0.7817      0.116      6.750      0.000       0.555       1.009
alcohol                  0.1028      0.031      3.293      0.001       0.042       0.164
========================================================================================
```

- Can predict 27% of the quality from the independent variables

- fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates, and alcohol are statistically significant.

White Wine Model with Predictions



Conclusion

Conclusion - How can you produce higher quality wine?

- Wine is complicated biological where yeast and bacteria convert sugars into alcohol. This process along with the aging process are meticulously monitored and enhanced to create quality wine.

Questions you might ask

- How can I produce higher quality wine without significantly altering my wine production?

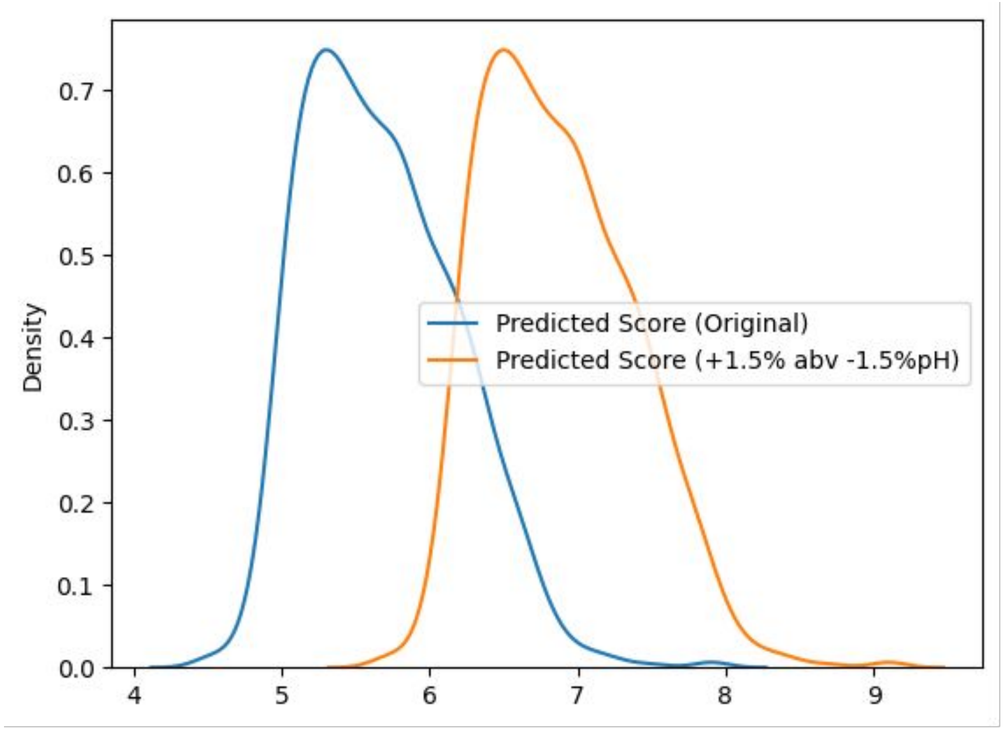
- How can I produce higher quality wine without significantly increasing production costs?

Conclusion

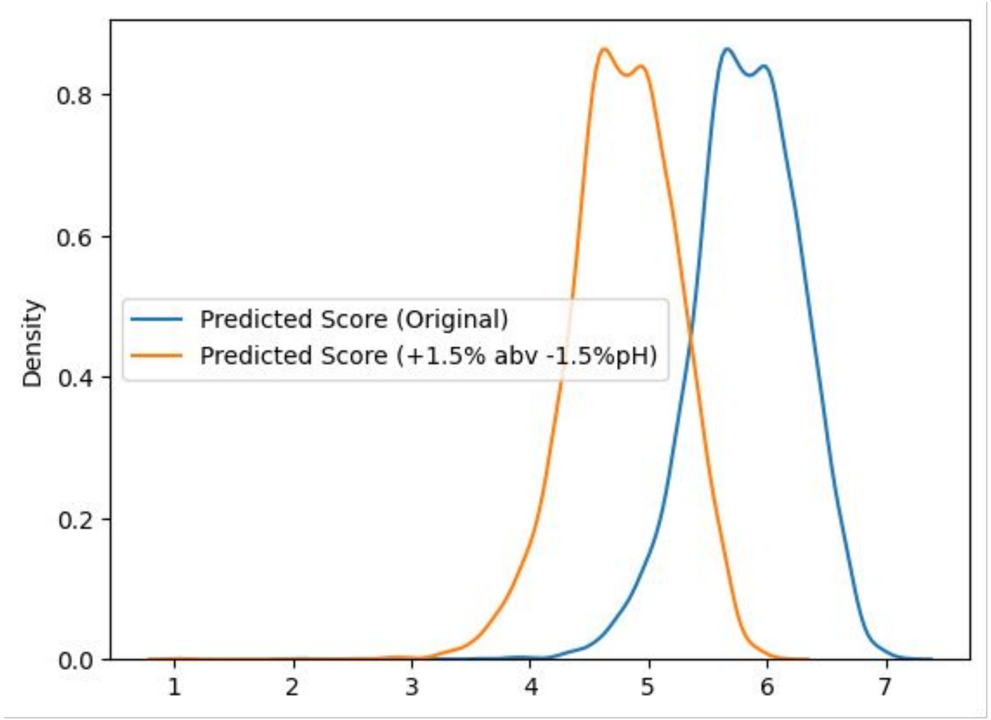
Our model identified two modifiable attributes post-fermentation: **' pH'** and **'alcohol'**.

Our model predicts winemakers can increase quality scores by **1 unit** by changing pH by 1.5 units and increasing alcohol by 1.5 units.

Red Wine Model with predicted increase



White Wine Model with predicted increase



Recommendation

- **1. Modify pH**
  - Decrease pH in Red Wine by adding Phosphoric acid to the finished wine prepacking
  - Increase pH in White Wine by adding carbonate salts prepackaging
- **2. Increase alcohol** by back-adding higher abv wine or other additive prepackaging

Bibliography

- https://archive.ics.uci.edu/dataset/186/wine+quality

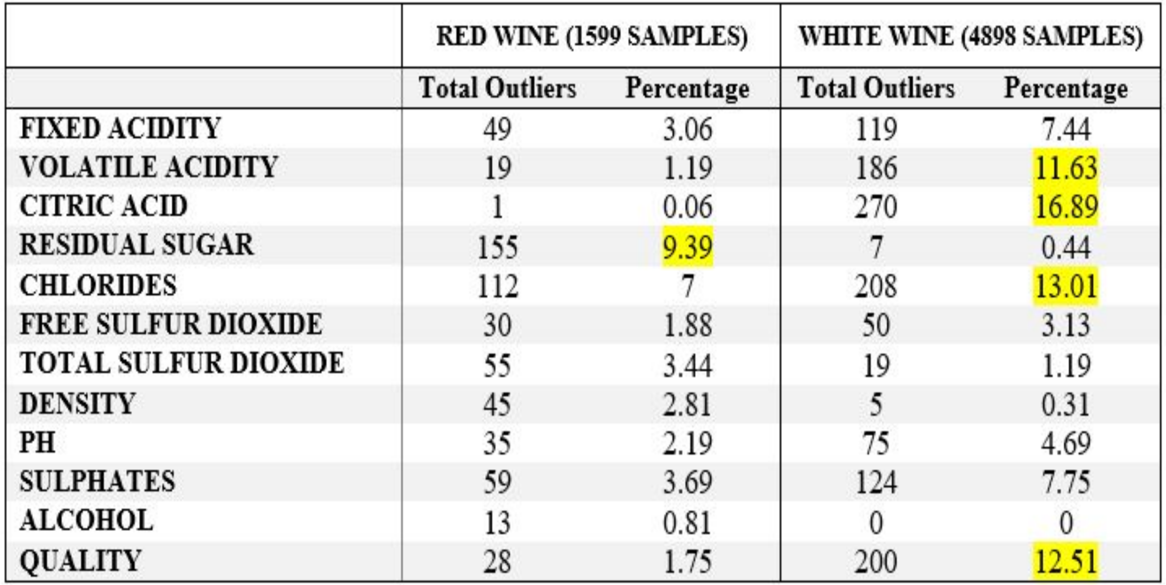
Backup Slides

Exploratory Data Analysis

Quantifying the outliers:

- **Moderate amount** of outliers in residual sugar for red wine.
- **High amount** of outliers for acid attributes, chlorides, and quality for white wine.

Outliers can negatively impact our statistical models.



	RED WINE (1599 SAMPLES)		WHITE WINE (4898 SAMPLES)	
	Total Outliers	Percentage	Total Outliers	Percentage
FIXED ACIDITY	49	3.06	119	7.44
VOLATILE ACIDITY	19	1.19	186	11.63
CITRIC ACID	1	0.06	270	16.89
RESIDUAL SUGAR	155	9.39	7	0.44
CHLORIDES	112	7	208	13.01
FREE SULFUR DIOXIDE	30	1.88	50	3.13
TOTAL SULFUR DIOXIDE	55	3.44	19	1.19
DENSITY	45	2.81	5	0.31
PH	35	2.19	75	4.69
SULPHATES	59	3.69	124	7.75
ALCOHOL	13	0.81	0	0
QUALITY	28	1.75	200	12.51

Total Outliers