# writeup

## STA 210 - Summer 2022

Team 2 - Alicia Gong, Ashley Chen, Abdel Shehata, Claire Tam

6-16-2022

### Introduction and Data

**Introduction** About 234 million hectoliters of wine were consumed in 2020, worldwide, with the US making up approximetly 14% of that consumption (Karlsson 2020). Since Wine composition and wine quality varies widely, it raises the question: what makes a good wine?

To answer that question, we will analyze the wine quality dataset from Vinho Verde vinyard in Portogal, and more importantly try to narrow down our question to make it possible for it to be supported by evidence. Below is the introduction to our research:

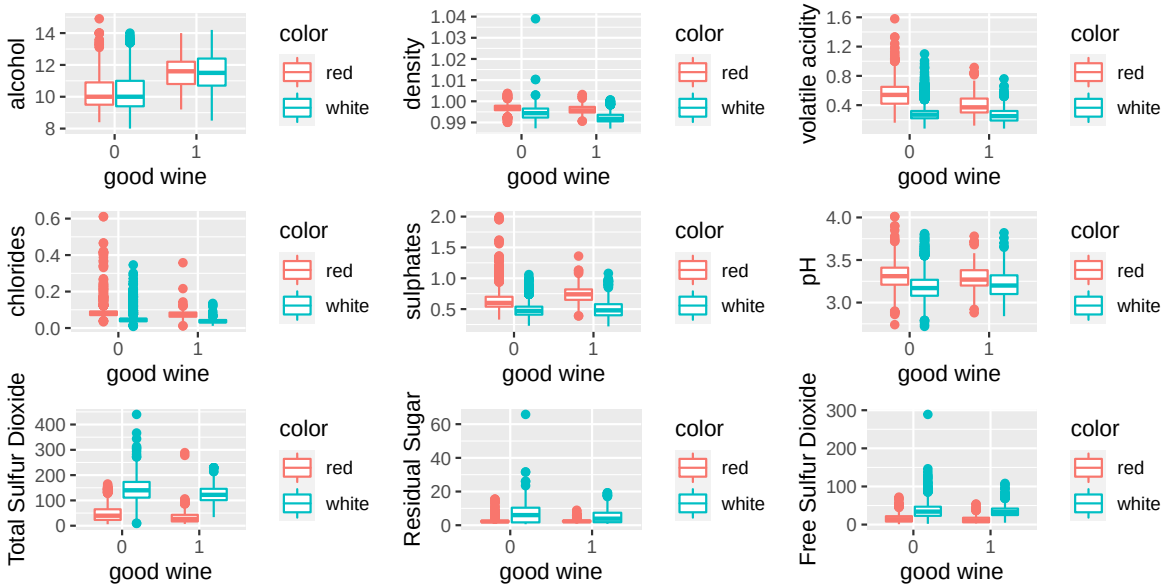
Project Goal: To identify variables that are important in explaining variation in the response. "Vinho Verde" is the kind of wine exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal. The Vinho Verde wine has its own unique system of production and is only produced from the indigenous grape varieties of the region. Vinho Verde region is one of the largest and oldest wine regions in the world, and is home to thousands of producers, generating a wealth of economic activity and jobs, and strongly contributes to the development of Minho province and the country. The Vinho Verde wine also enjoys high reputation worldwide. It is recurrently awarded in national and international competitions. The goal of this dataset is to model wine quality based on physicochemical tests. We believe that this dataset can also be used to analyze the relationship between different chemical compositions and the ratings of wine quality. We believe that this dataset can also be used to analyze what chemical factors are attributable to the final rating of Vinho Verde wine. Our research may shed light on future research and development directions for improving the quality of Vinho Verde wine, which may also contribute to the competitiveness of Portuguese wine industry.

Our goal is to produce a classification model that best explains how different chemical compositions of the Portuguese "Vinho Verde" wine affects the variation of the wine quality.

**Data Introduction** The Wine Quality dataset was collected from Vinho Verde wine Samples, from the North of Portugal. The data was originally donated in 2009 by Professor Cortez. The specific mechanism of the collection of the data was lab work done on different wines to measure their chemical attributites (like acidity etc.). The quality of the wine however was obtained through the average rating of three wine experts. The dataset is divided into two: Red wine and White wine. Red Wine has 1599 observations, and white wine has 4898 observations (each observation being a specific wine). Information about the wine include but are not limited to:PH,Density,Acidity, and alcohol content.

Each observation is a specific wine from the Vinho Verde region. Thus, there might be a little uncertainity in collecting the exact numbers for each numbers. However, since the Vinho Verde region is a vast region spreading 15500 hectareas of vineyards in far-north Portugal, this uncertainity shouldn't be significant in our analysis or project. Thus, we will assume that the datas are independent and random


**Exploratory Data Analysis**




## Methodology

**Best AIC Model** We choose best AIC model. What is that? Why? Reason and justification.

Data split: we split the data into 25% testing set and 75% training set. ::: {.cell}

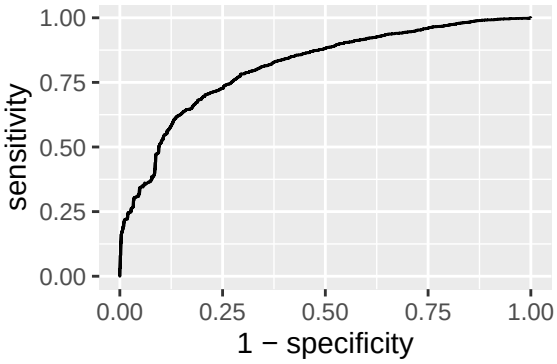
:::

## Model output

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 406.893 | 78.634 | 5.175 | 0.000 |
| c_fixed.acidity | 0.465 | 0.077 | 6.005 | 0.000 |
| c_volatile.acidity | -3.271 | 0.420 | -7.791 | 0.000 |
| c_residual.sugar | 0.216 | 0.031 | 6.941 | 0.000 |
| c_chlorides | -7.312 | 2.836 | -2.578 | 0.010 |
| c_free.sulfur.dioxide | 0.012 | 0.003 | 3.578 | 0.000 |
| c_total.sulfur.dioxide | -0.005 | 0.002 | -3.065 | 0.002 |
| c_density | -427.460 | 79.687 | -5.364 | 0.000 |
| c_pH | 2.474 | 0.422 | 5.861 | 0.000 |
| c_sulphates | 2.665 | 0.327 | 8.158 | 0.000 |
| c_alcohol | 0.440 | 0.096 | 4.583 | 0.000 |
| colorwhite | -0.592 | 0.283 | -2.093 | 0.036 |

### ROC curve prediction

```
New names:
* `` -> ...1
```

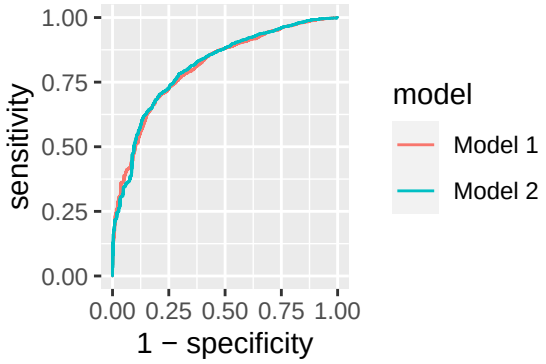



### Model Evaluation

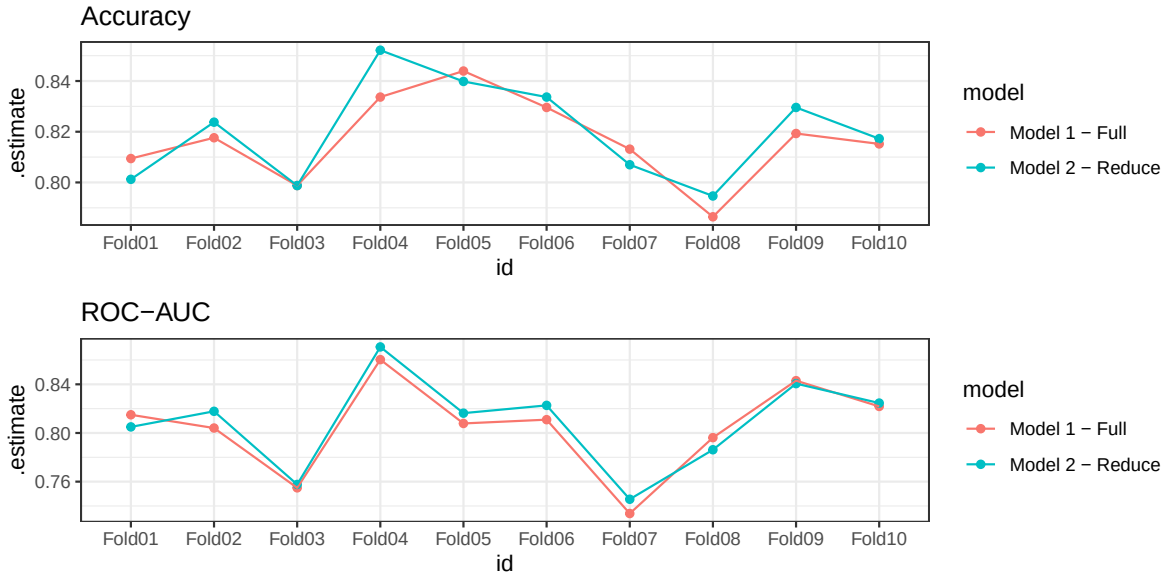
**Logistic Model** We compare the best AIC model with a logistic model with interactive terms.

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -2.151 | 77.953 | -0.028 | 0.978 |
| c_fixed.acidity | 0.075 | 0.052 | 1.433 | 0.152 |
| c_volatile.acidity | -3.554 | 0.429 | -8.276 | 0.000 |
| c_residual.sugar | 0.141 | 0.025 | 5.586 | 0.000 |
| c_chlorides | -12.100 | 3.041 | -3.980 | 0.000 |
| c_total.sulfur.dioxide | -0.014 | 0.004 | -3.496 | 0.000 |
| c_density | -10.347 | 77.813 | -0.133 | 0.894 |
| c_sulphates | 3.583 | 0.578 | 6.196 | 0.000 |
| c_alcohol | 0.944 | 0.126 | 7.470 | 0.000 |
| color_white | 265.024 | 79.784 | 3.322 | 0.001 |
| c_total.sulfur.dioxide_x_color_white | 0.015 | 0.004 | 3.524 | 0.000 |
| c_density_x_color_white | -262.834 | 79.290 | -3.315 | 0.001 |
| c_sulphates_x_color_white | -1.469 | 0.690 | -2.128 | 0.033 |
| c_alcohol_x_color_white | -0.374 | 0.145 | -2.575 | 0.010 |

**Model Selection** The two ROC curves are very close, suggesting that the two models' performances are similar.




| model | .metric | .estimator | .estimate |
|---|---|---|---|
| Logistic Model | roc_auc | binary | 0.809 |
| Best AIC Model | roc_auc | binary | 0.812 |




**Cross Validation** To perform the cross validation, we split the training data into 10 folds. The two models have similar accuracy and the roc-auc for 10 folds, so due to the principles of parsimonious, we prefer the reduce one.

**Check Conditions** When conducting a logistic regression, it is important to check the conditions are satisfied. When evaluating if the log-odds have a linear relationship with the predictors, we find that the conditions for linearity are not satisfied for all the variables. The linearity conditions are particularly not fulfilled for the variables free sulfur dioxide (the spread takes on a curved shape), total sulfur dioxide (the spread has similarly curved shape), and sulphates (the data points appear to be randomly dispersed).



The independence conditions, however, are satisfied. The Vinho Verde region is a vast region spreading 15500 hectareas of vineyards in far-north Portugal allowing the observations to be collected independently from each other and the accumulated data are not collected across an extended period of time.

**Results**

- full model
- best AIC model
- AIC BIC ROC
- model selection: do not include the interactive terms due to parsimonious. because the ROC does not improve much.

From our analysis, we found that the variables that contribute significantly to wine quality were fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, and sulfates. Beta weights can be rank ordered to determine which variable is the strongest predictor. Variables with the biggest beta have the strongest influence in wine quality which can be determined from the regression coefficient. The predictors from strongest to weakest influence in wine quality were

## Discussion & Conclusion

To answer our research question "what chemical components contribute to the quality of wine?" we implemented a stepwise AIC test to determine model with the most significant predictor variables and calculated the beta weights and p-values for each of the variables.

- what chemical components contribute to wine quality
- future research suggestions

In the future, to improve the accuracy of the model, different stepwise algorithms (p-values for example) can be evaluated and adjusted. There are many different methods such as through feature engineering, viewing potential interaction terms, or other performance measurements such as machine learning algorithms to better predict our results.

- suggestions to wine valley?

## Reference

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009.