# Modeling wine preferences by data mining from physicochemical properties

**Bob Ding, Lynn Fan, Alice Jiang**

**Dec. 2nd, 2018**

## What Makes a Good Glass of Wine

Project Goal: **Explanation**.
To identify variables that are important in explaining variation in the response.

We are interested in **researching what factors contribute to the quality of wine for different types of red vinho verde from Portugal**

What makes good glass of wine? How do wine experts evaluate whether a wine satiate human palettes? We picked red vinho verde from Portugal to conduct our research. The data set in this research was used to predict quality of wine for future wine certification, complementary to human wine tasters, in the paper we cited. We believe that this dataset can also be used to analyze what chemical factors are attributable to the final rating of wine. Understanding what makes a good wine may shed light on future directions for chemical methods that could improve/preserve wine quality.

## Why Proportional Odds & Logistic Model

**Proportional Odds**
for the professionals

**Proportional odds model** gives more specific quality level interpretation, which is more informative for professionals in the wine industry.

**Logistic regression** indicates whether the wine is 'good' or 'bad'. For the potential customers, simple and straightforward information is more useful.

**Logistic Regression**
for the commons

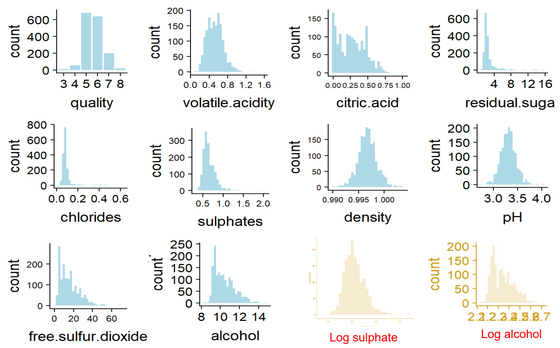
## Data Explanation & Data Exploration

**Response Variable:**

- **quality:** the quality of the wine (a score between 0 and 10)

**Explanatory Variables:**

- **fixed.acidity:** the amount of acid in wine that's not volatile (do not evaporate fast)
- **volatile.acidity:** the amount of acetic acid in wine
- **citric.acid:** citric acid is found in small quantities, and can add freshness and flavor to wines
- **residual.sugar:** the amount of sugar left after fermatation stops
- **chlorides:** the amount of salt in the wine
- **free.sulfure.dioxide:** the free form of S02S02 exists in equilibrium between molecular S02S02 (as a dissolved gas) and bisulfite ion
- **total.sulfur.dioxide:** amount of free and bound forms of S02S02
- **density:** the density of the liquid
- **pH:** the indicator of the acidicity or basic property of the wine
- **sulphates:** a wine additive which can contribute to sulfur dioxide gas S02S02 levels
- **alcohol:** the percent alcohol content of the wine




- Many variables are normally distributed, which might be dominating for future logistic analysis
- residual.sugar, cholorides, and sulphates are slightly skewed to the right
- free.sulfur.dioxide and alcohol have an obvious rightly skewed distribution.
- citric.acid at first appears to have a bimodal distribution, because there are some wines with zero citric acid, Based on data definition, we know it is possible for wines to have citric acid of 0.
- The variables in yellow plots are logged sulphate, and logged alcohol; log transformation is used to improve model fit by normalizing data

## Conclusion

- Our project goal is explanation: to identify the chemical factors that make significant contribution to the quality of wine
- As quality is a discrete variable, we used logistic model and ordinal logistic model for our research, serving customers and wine professionals, respectively
- Two models congruently leads to the conclusion, that for vinho verde wine produced in Portugal, **alcohol level**, **acetic acid level**, and **sulfate level** are very strong explaining factors
- citric acid level and free sulfur dioxide level contribute to the variation in quality, though not as influential

**Alcohol** **Volatile acetic acid** **Sulfate**

## Limitation

- **Extrapolation**: the sample data has wine with quality mostly ranging from 4 to 7. For future uses, it is very important not to extrapolate beyond this quality range, or the model will not give valuable explanation
- **Assumption concern:** In both models, there are some slightly non-random patterns for the binned residual plots for a few variables, which require care and possible future analysis – we did not conduct further improvements as they are beyond our knowledge learned in class
- **Test set:** Model should also be tested on larger and more comprehensive datasets and more chemical factors can be included for future modeling

## Ordinal Regression Model Detail

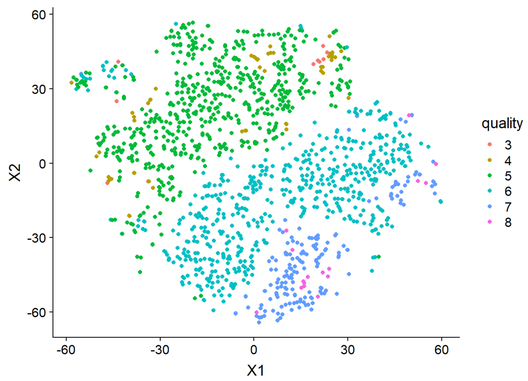
| term | estimate | std.error | statistic | coefficient_type |
|---|---|---|---|---|
| volatile.acidityCent | -3.60 | 0.38 | -9.53 | coefficient |
| citric.acid | -0.74 | 0.38 | -1.97 | coefficient |
| chloridesCent | -5.26 | 1.31 | -4.02 | coefficient |
| sulphatesCent | 2.73 | 0.36 | 7.55 | coefficient |
| pHCent | -1.48 | 0.41 | -3.56 | coefficient |
| alcoholCent | 0.95 | 0.06 | 16.37 | coefficient |
| 3\|4 | -6.08 | 0.34 | -17.73 | zeta |
| 4\|5 | -4.14 | 0.18 | -23.23 | zeta |
| 5\|6 | -0.50 | 0.12 | -4.27 | zeta |
| 6\|7 | 2.32 | 0.14 | 16.97 | zeta |
| 7\|8 | 5.33 | 0.27 | 19.74 | zeta |

- **ChloridesCent**, **Volatile.acidity**, **sulphatesCent** are very significant and influential predictors
- Model meets all assumptions in statistics are big and p-values are small

## Ordinal Regression Model Output

| pred.comp | Accuracy | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 4 | 0.5 | 1 | NA | NA | NA | NA | NA |
| 5 | 0.72 | 8 | 39 | 510 | 223 | 9 | NA |
| 6 | 0.63 | 1 | 14 | 167 | 386 | 138 | 10 |
| 7 | 0.62 | NA | NA | 3 | 29 | 52 | 8 |
| 8 | 0.5 | NA | NA | 1 | NA | NA | NA |

- Model not predicting quality = 1, 2, 3, 9,10 as no (few) raw data available
- Model not predicting well at quality = 4, 8 because of dominance
- Model predicting well at quality = 5, 6, 7, as bigger data size available
- Overall accuracy: 59.47%; Class 5 accuracy: 72.23%




- Data dominance in class 5, 6. This explains why prediction at 5, 6 is better
- Overall, the dataset is a good dataset for the ordinary regression model, as the data structure suggests

## Ordinal Regression Model Inference

- **Alcohol level**, **acetic acid level**, and **sulfate level** are the strongest explaining variables for quality with largest magnitude of test statistic.
- During modeling process, we tried to log transform sulfate level as we discovered nonrandom patterns in binned residual plot for sulphatesCent, but they still exist after log transformations. We believe that the patterns are partly caused by mid-quality wines dominating the model.
- The model makes valuable predictions for mid-quality wines as the accuracy rate exceeds 70% for quality=5, 6 and 7. As they are the most common occurring quality for vinho verde, professionals might find this model useful.

**Alcohol** **Volatile acetic acid** **Sulfate**

*References:*

*Vinho Verde. (2018, September 20). Retrieved from https://en.wikipedia.org/wiki/Vinho_Verde*

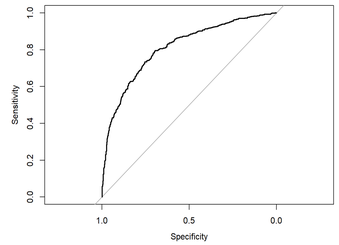
*Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems,47(4), 547-553. doi:10.1016/j.dss.2009.05.016*

## Logistic Regression Model Detail

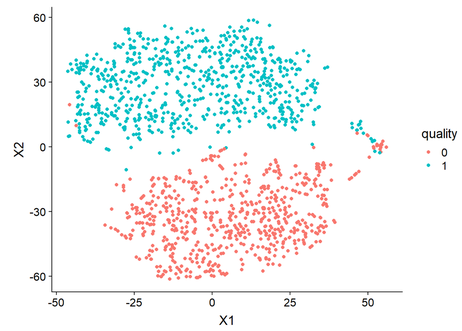
| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 1.90 | 0.23 | 8.30 | 0.00 |
| alcoholCent | 1.06 | 0.08 | 13.54 | 0.00 |
| volatile.acidityCent | -3.28 | 0.47 | -7.00 | 0.00 |
| log_sulphates | 2.71 | 0.36 | 7.59 | 0.00 |
| chloridesCent | -2.51 | 1.53 | -1.64 | 0.10 |
| citric.acid | -1.35 | 0.45 | -2.98 | 0.00 |
| pHCent | 0.97 | 0.86 | 1.13 | 0.26 |
| free.sulfur.dioxideCent | -0.04 | 0.01 | -3.07 | 0.00 |
| log_sulphates:pHCent | 5.12 | 1.74 | 2.94 | 0.00 |
| alcoholCent:free.sulfur.dioxideCent | 0.03 | 0.01 | 3.60 | 0.00 |
| log_sulphates:free.sulfur.dioxideCent | -0.08 | 0.03 | -2.95 | 0.00 |

- **Volatile.acidityCent**, **log_sulphates**, **log_sulphates:pHCent**, and **chloridesCent**, and are most influential predictors.
- **pHCent** and **chloridesCent** have p-values exceeding the 0.05 threshold, so the extent of their impact is not significant.
- The remaining variables are very strong predictors with p-value = 0, but the magnitude of impact is not that large as their coefficients are relatively small.
- We believe that acidity

## Logistic Regression Model Output




- From the ROC curve and AUC calculation, we can see the curve is close to the top left corner, area= 0.81.
- This shows that the logistic model can distinguish between good and not good quality, so this is a pretty good model.




- The t-SNE plot shows that the logistic regression the raw dataset has been a good model in differentiating good and bad wine

## Logistic Regression Model Inference

- **Alcohol** level indicates fermentation process. So it has positive effect.
- **Sulfates** help preserve freshness. So it has positive effect.
- **Volatile acetic acid** leads to a sour taste of vinegar, so it has a negative coefficient.
- The assumptions are met except a few observations with high leverage, but they have low Cook's distance values, so they do not have a big impact on the predicting power of model.

**Alcohol** **Volatile acetic acid** **Sulfate**