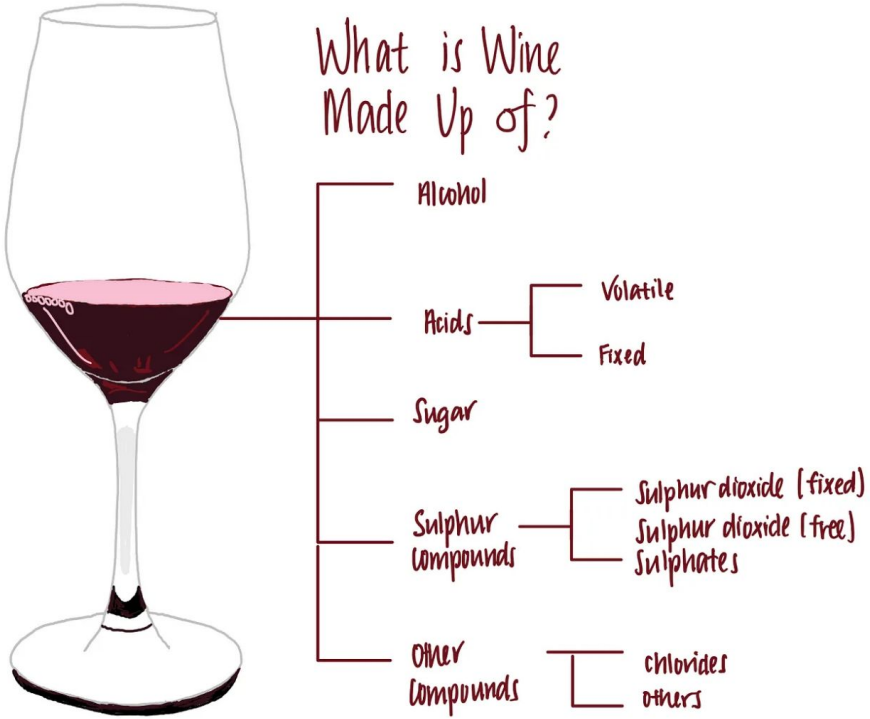
ANALYSIS OF

WINE QUALITY

Nguyen Dong, Ian Mccarthy, Nan Wang, Xinyu Wang, Jessica Xu

Jessica Xu, Zhengyi Zhang

Introduction/Background



Nowadays,The quality of wine is primarily assessed through physicochemical and sensory evaluations.

This dataset, which shows red wines quality and their chemical properties, is sourced from the Vinho Verde region in Northern Portugal

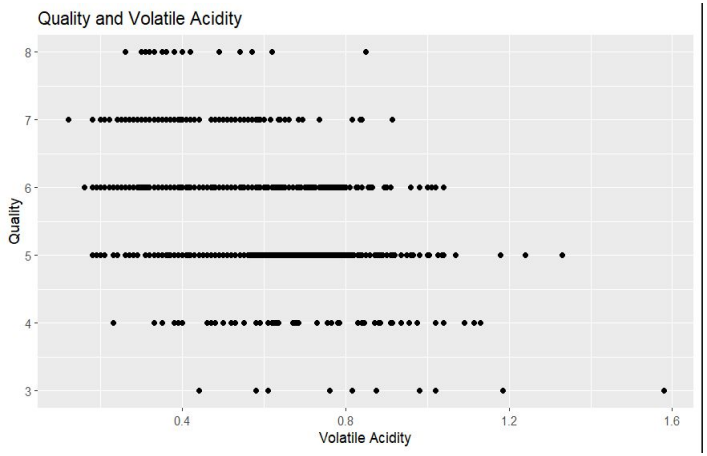
The goal is to analysis a model that forecasts wine quality from physicochemical tests, exploring how factors correlate with the wine's quality.

Refs: Tang, Travis. “What Makes Great Wine... Great?” *Medium*, Towards Data Science, 28 Oct. 2020

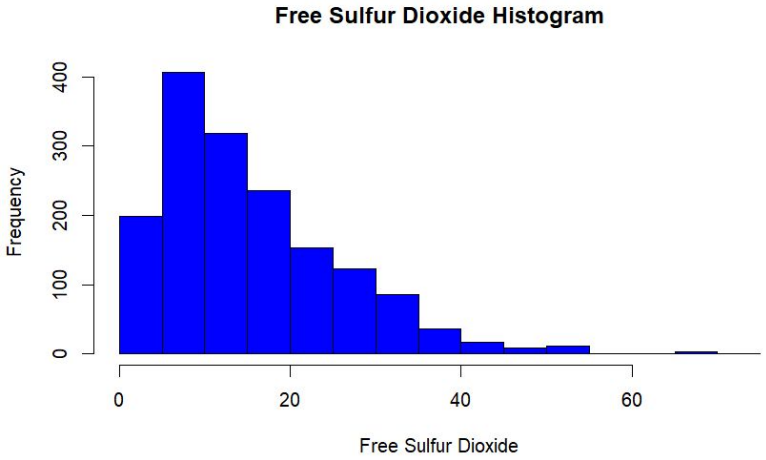
Variables In the Red Wine Quality Dataset

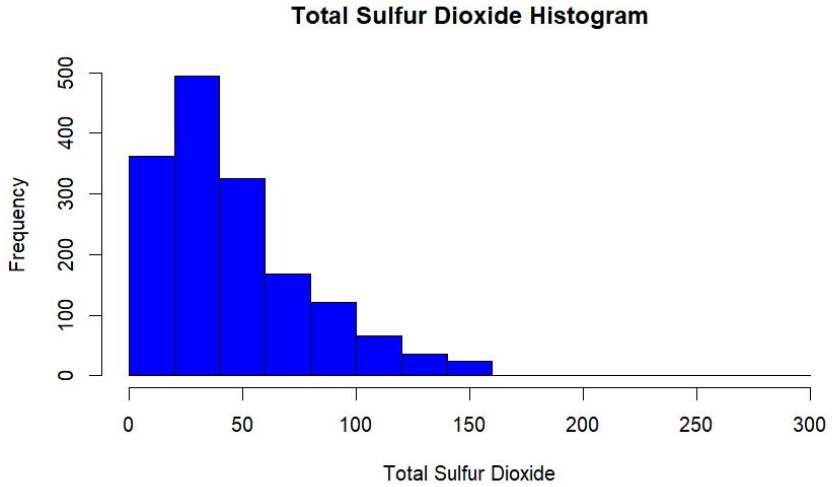
Variable Name	Description
Fixed Acidity	Concentration of tartaric acid (g/dm^3)
Volatile Acidity	Amount of acetic acid (g/dm^3)
Citric Acid	Citric acid content (g/dm^3)
Residual Sugar	Remaining sugar after fermentation (g/dm^3)
Chlorides	Amount of salt (g/dm^3)
Free Sulfur Dioxide	Free form of SO2 (mg/dm^3)
Total Sulfur Dioxide	Total SO2 content (mg/dm^3)
Density	Density of wine (g/cm^3)
pH	Acidity/basicity level
Sulphates	Sulphates content (g/dm^3)
Alcohol	Alcohol percentage (%)
Quality	Quality score (0-10)

What data looks like



Histograms of some of the x’s





Correlation plot and VIF Values

VIF:

X1 = 7.77 X2 = 1.79

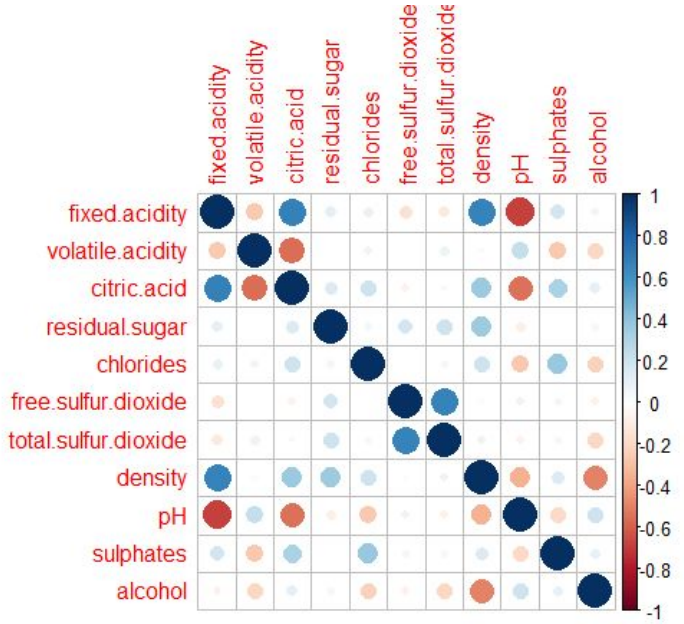
X3 = 3.13 X4 = 1.70

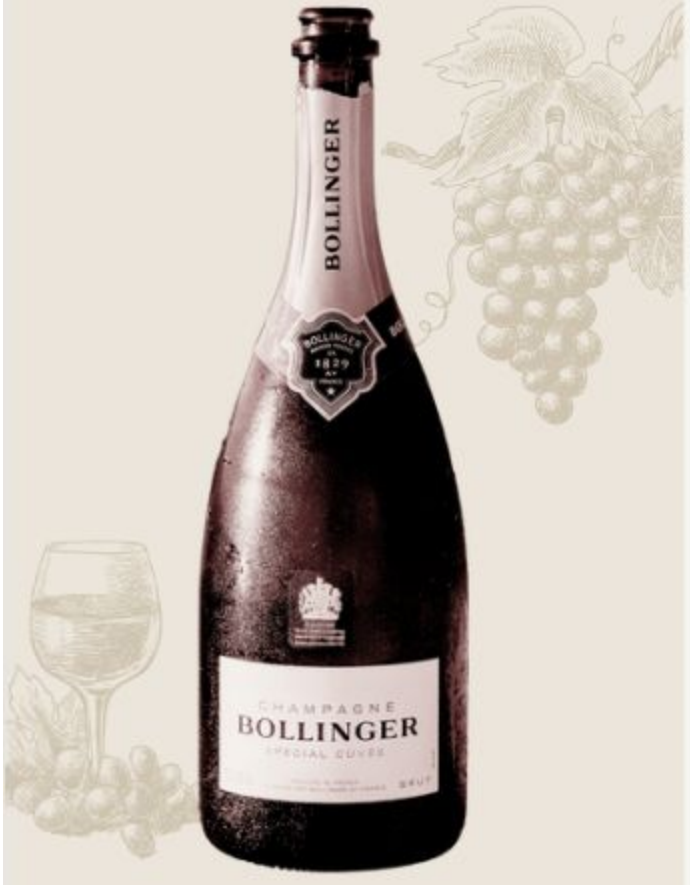
X5 = 1.48 X6 = 1.96

X7 = 2.19 X8 = 6.34

X9 = 3.33 X10 = 1.43

X11= 3.03





BEST MODEL

$$Y = 5.649 - 0.581x2 - 0.494x5 + 0.179x6 - 0.303x7 - 0.171x9 + 1.022x10$$

STEP 1

Square root the citric acid variable and Log transformation on all the rest of the variables except residual sugar, density, and pH due to right skew (the three variables mentioned didn't have a right skew)

STEP 2

Conduct Forward Backward model selection with BIC criteria

STEP 3

Remove the the alcohol variable from that model since it had a VIF of 3.3

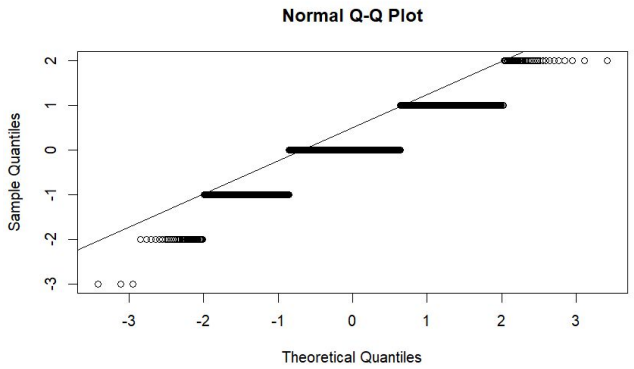
STEP 4

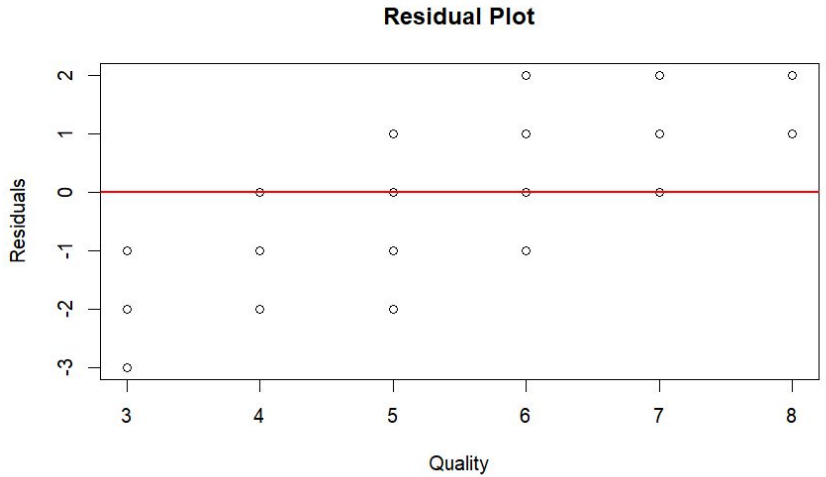
Round the fitted values of the model to the nearest whole number (round down if below 0.5, round up if 0.5 or greater) and recalculate the residuals

Model graphs/normality

Shapiro: p < .001

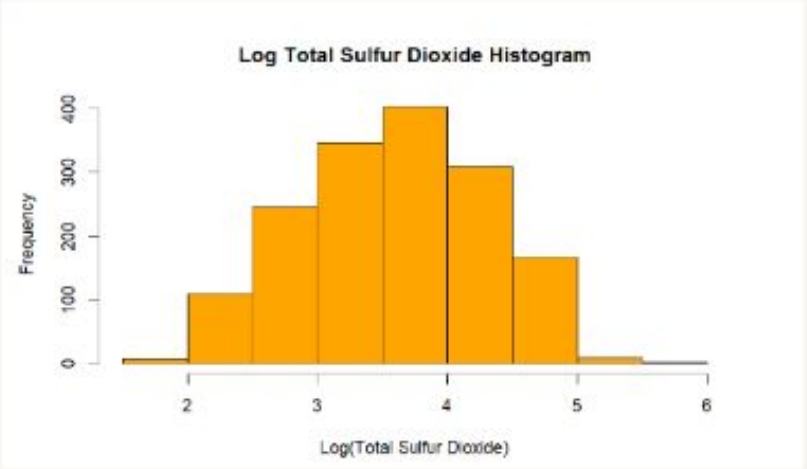
FK: p = 0.5405

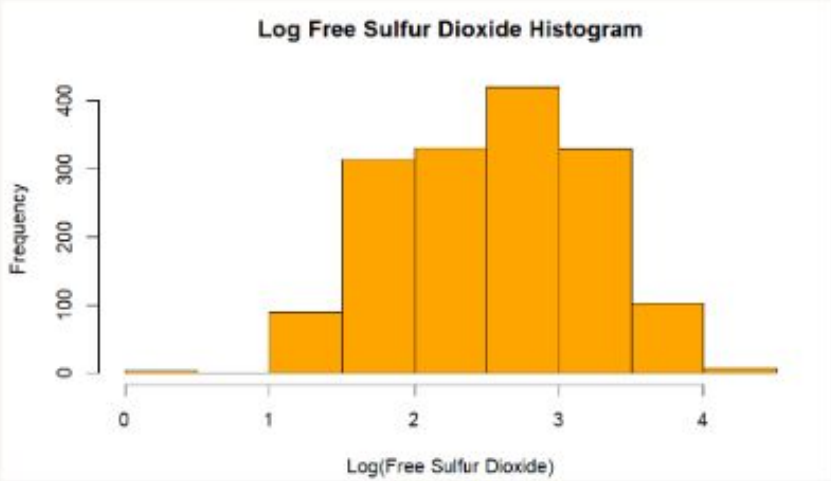




HISTOGRAM

AFTER TRANSFORMATION





=> More evenly distributed, stabilize the variance, make the data more symmetric for better model

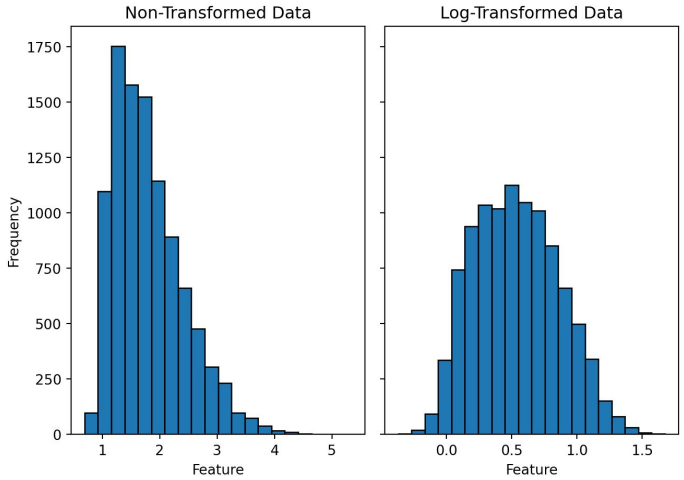
Issues with the data

Quality is categorical, so poses an inherent issue with a model that outputs continuous fitted values Continuous values imply a level of precision or magnitude that may not align well with the discrete nature of categorical variables.

Issues with our model

Even if assumptions of constant error variance and independent observations are met, normality still not hold, the reason probably is the complex relationships among variables.

Transforming the independent variables with a logarithmic function complicates their interpretation, as it alters the scale of the data and may require additional effort to understand the relationships between the variables.



Possible steps to further explore

1. Trying modeling with something that adapts to discrete outcome variables more efficiently
2. Investigate alternative variable transformations or scaling methods that preserve interpretability while addressing issues related to variable distribution.