Basics

Dataset: https://archive.ics.uci.edu/dataset/186/wine+quality
Repository: https://github.com/iamhex7/151A-Wine-Quality.git
Research Question:
How do chemical properties differently affect the quality of white and red wines?

0. Data and Modeling Objective

The original data comes in two datasets: one contains 4,898 white-wine observations and another one with 1,599 red-wine observations.
We manually merged them two into a single dataframe with one extra categorical variable called type indicating wine type – 0 as red wind, 1 as white wine.
Response Variable: quality (integer 0–10)
Key grouping: type (red / white as dummy variable)
(Potential) Explanatory Variables: fixed_acidity, volatile_acidity, citric_acid, residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH, sulphates, alcohol.

1. determining if a transform is needed

Plot for each continuous x:

- histogram / density plot
- summary statistics (mean, median, IQR, max/min)

Transform if:

- Spotted strongly right-skewed (long-tailed) pattern → transform: log(x) or sqrt(x)
- If a variable has a value of 0 and cannot use log, consider log(x + c) (c is a small constant such as 0.1).

2. Basic Main-Effects Model excluding interaction

```
m_main <- lm(quality ~ type + x1 + x2 + ... + x11, data = wine_transformed)
```

2.1 Check the residuals

- Residuals vs Fitted: look for non-linearity or heteroscedasticity?
- QQ plot: Is the error roughly close to normal?

2.2 Checking collinearity

- Calculate the correlation coefficient matrix for all x
- Calculate the VIF (variance inflation factors) on this model

If some variables are highly correlated:
If both variables are meaningful for the scientific question, they can be retained initially, and simplified later during model selection;

2.3 Which Features May Be Removed:

Examine the estimated value and standard error of each main effect

Examine their t-values/p-values

Compare the changes in AIC/BIC before and after removing a certain variable

3. Explore & test interaction: type × chemical

Chemically, below are likely strong interactions among features:

free sulfur dioxide & total sulfur dioxide

fixed acidity & pH

alcohol & density

residual sugar & density

citric acid & pH

volatile acidity & pH

3.1 Visual Review First, EDA

Are the slopes of the two lines significantly different for red wine and white wine?

For example:

```
ggplot(wine, aes(x = alcohol, y = quality, type = type)) +
geom_point(alpha = 0.2) +
geom_smooth(method = "lm", se = FALSE)
```

If the regression lines of the two colors are significantly non-parallel, then ⇒ strong candidate: type × alcohol interaction.

3.2 First, create a "full interaction" model and perform a global test.

```
m_int_full <- lm(quality ~ type * (x1 + x2 + … + x11), data = wine_transformed)
```

type * (…) :

All main effects (type + x1 + … + x11)

Add all type: xj interactions

Then:

Compare using ANOVA incremental F-test: anova(m_main, m_int_full)

Compare AIC / BIC: AIC(m_main, m_int_full) / BIC(m_main, m_int_full)

4. feature selection on the main effects

We have a model:

Some main effects: type + selected xj

Some interactions: type: xj

4.1 General Principles

Main effects corresponding to retained interactions should not be removed. Variables that do not appear in interactions and have small and unstable coefficients can be considered for removal.

4.2 Model Selection Procedure

- AIC/BIC
- Cross-validation

5. Model Diagnosis

- Residual vs Fitted: Check for nonlinearity and heteroscedasticity

QQ plot: Check for heavy tails

- Leverage & Cook's distance: Find high leverage/influential points