plan

RQ

How do key chemical attribute affect wine quality, and do these relationships differ between red and white wines in ways that are relevant to winemaking decisions?

这个版本最符合：- project instruction - 统计建模要求（interactions, linear modeling theory）- 教授的三个主要批评（lack of context、too broad、motivation unclear）

阶段 0：EDA

- 检查各变量分布，识别 skewness 并决定是否需要 transformation。
- 检查变量间的相关结构，识别 potential multicollinearity。
- 比较 red 与 white 的化学特征分布差异，识别 potential interaction。
- 根据 scatterplots 判断线性关系的方向与强弱。

此阶段不进行推断，不删除变量或观测点。

阶段 1：构建 Candidate Models

- 根据 EDA 决定必要的 transformations（如 log-transform）。
- 删除无意义或高度共线的变量（如 density）。
- 依据科学动机与 EDA 结果决定主要交互项：strong、moderate、weak。
- 构建满足 marginality 的 hierarchical models。

本阶段只定义候选模型，不做检验或推断。

阶段 2：拟合模型

拟合多个 candidate models 以便后续比较，例如：

```
mA <- lm(...)
mB <- lm(...)
mC <- lm(...)
mD <- lm(...)
```

阶段 3：模型选择（Model Selection）

允许使用：- AIC、BIC - Nested model F-tests（ANOVA） - Adjusted R^2 - 科学解释性与可解释性

不允许使用：- stepwise regression - 使用 AIC/BIC 选择模型后再做未修正的 classical inference

原则： - AIC/BIC 仅用于在 candidate models 间选择。 - 若差异显著，选择 AIC/BIC 较小的模型；若差异不显著，选择更简单、解释性更强的模型。 - 最终模型确定后才能做 p-values、t-tests、CI 等推断。

阶段 4：模型诊断（Model Diagnostics）

检查四类假设：

1. Linearity
   工具：Residuals vs Fitted
   如出现曲线趋势，考虑 transformation 或添加交互项。
2. Constant variance（Homoscedasticity）
   工具：Residuals vs Fitted
   若异方差明显，可进行 predictor transform 或在结论中说明限制。
3. Normality of Errors
   工具：QQ plot
   大样本下，轻微非正态不影响估计量的近似正态性（由 CLT 保证）。
4. Influential Points
   工具：Cook's Distance、Residuals vs Leverage
   非数据录入错误的观测不可删除；仅在显著影响模型时记录其影响。

本阶段不修改数据，除非发现确凿的数据错误。

阶段 5：是否需要 L1 / L2 回归（可选）

- L1（LASSO）与 L2（Ridge）不能用于最终推断模型。
- 可作为 sensitivity analysis 检查系数稳定性。
- 在存在 multicollinearity 或系数不稳定时可使用。
- 典型适用情形：free_sulfur_dioxide 与 total_sulfur_dioxide、residual_sugar 与 density。

结果仅作为补充，不进入最终推断框架。

阶段 6：最终模型推断（Inference）

在最终模型确定且 diagnostics 通过后：

- 检查 main effects 的方向、大小、显著性。
- 解释 interaction effects，阐述红白葡萄酒之间的差异。
- 绘制 interaction plots 说明斜率差异。
- 将统计结果与酿酒背景（chemistry/winemaking decisions）联系。

此阶段给出最终回答，即回答 research question。

阶段 7：写结论（Conclusion）

包含内容：

- 最终模型的主要科学发现（main effects + interactions）。
- red 与 white wine 之间的系统性差异。
- 哪些化学属性最重要、如何影响 quality。
- 交互项揭示的 winemaking insights。
- 模型局限性：heteroscedasticity、non-normal tails、limited R^2、未观测变量等。