**Data Source and Description**

The dataset used in this analysis is the *Red Wine Quality* dataset, which contains physicochemical measurements (numerical variables) of red Portuguese “Vinho Verde” wine samples, along with a quality score (integer between 0 and 10) assigned by wine tasters. The dataset includes **1,599 observations** and **12 variables**, of which **11** are potential predictors and **1** is the response variable.

**Variables:**

Dependent/response variable: quality

Independent/predictor variables: fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol

### ****Data Preparation****

The dataset was read into R and checked for missing values, incorrect data types, and outliers.  
All variables were treated as continuous except for the response variable quality, which was modelled as a numeric score.

If required, data transformations (e.g., log or square-root) were considered to reduce skewness or stabilize variance in predictors.

**Model Specification**

The relationship between wine quality and the physicochemical predictors was modelled using **multiple linear regression**:

qualityi = β0 + β1 x1i + β2 x2i + ⋯ + βp xpi + εi,

where x1i, … , xpi​ are the physicochemical properties of wine i,  
βj are regression coefficients, and εi are random errors assumed to be independent and normally distributed with mean 0 and constant variance σ2.

The full model included all 11 predictors:

### ****Model Selection****

To identify a parsimonious model that adequately explains wine quality, **stepwise selection** based on the Akaike Information Criterion (AIC) was applied using both forward and backward selection:

The final selected model retained only significant predictors (p < 0.05) and achieved a balance between explanatory power and simplicity.

### ****Model Diagnostics****

Model assumptions were checked through residual analysis:

1. **Linearity:** Scatterplots of residuals versus fitted values.
2. **Normality:** Q–Q plot of standardized residuals.
3. **Homoscedasticity:** Breusch–Pagan test and residual plots.
4. **Multicollinearity:** Variance Inflation Factors (VIF).

### ****Interpretation and Prediction****

Coefficients from the final model were interpreted to assess how each physicochemical property influences wine quality. Model performance was summarized using R2, adjusted R2, residual standard error, and AIC.  
Finally, predicted values were compared with observed quality scores to evaluate model fit.