Úkol A – Průzkumová analýza dat a lineární regrese: kvalita bílého vína

# Úvod

Cílem této analýzy je prozkoumat dataset obsahující informace o bílém vínu z oblasti Minho v Portugalsku. Každý vzorek vína byl hodnocen degustátory na škále od 0 do 10 a zároveň u něj bylo změřeno 11 fyzikálně-chemických parametrů. Cílem je určit, které vlastnosti vína ovlivňují jeho subjektivní hodnocení, a vytvořit lineární regresní model predikující kvalitu vína na základě těchto parametrů.

# Metody

Pro analýzu byly použity nástroje jazyka R – základní funkce (boxplot, hist, lm, summary), dále knihovny ggplot2, dplyr a corrplot. Analýza probíhala ve dvou částech:  
  
1. Průzkumová analýza dat : kontrola dimenzí, chybějících a odlehlých hodnot, distribuce proměnných, transformace a vizualizace vztahů.  
2. Lineární regrese: vytvoření modelu, interpretace koeficientů, hodnocení přesnosti a statistické významnosti.

# Výsledky

## Dimenze a typy dat

Dataset obsahuje 4898 řádků (pozorování) a 13 sloupců. První sloupec (ID) nebyl dále analyzován. Zbývajících 12 sloupců obsahuje fyzikálně-chemické vlastnosti a senzorické hodnocení. Proměnná sweet je binární (0 = suché, 1 = sladké), proměnná quality je ordinální a slouží jako cílová veličina.

## Chybějící a odlehlé hodnoty

Pomocí is.na() bylo ověřeno, že v datasetu se nevyskytují žádné chybějící hodnoty. Boxploty ukázaly odlehlé hodnoty např. u sulphates a total sulfur dioxide, ale tyto byly ponechány, protože odpovídají reálným fyzikálním rozsahům. Jediný, který neobsahuje odlehlé hodnoty, je sloupec s hodnotami alkoholu.

## Transformace dat

Proměnné chlorides, volatile acidity a citric acid měly výrazně nesymetrické rozdělení. Pro zajištění normality byly transformovány pomocí logaritmu: log(x + 1).

## Korelační analýza

• Nejvyšší negativní korelace: alcohol vs. density (r = −0.78)  
• Nejvyšší pozitivní korelace: free sulfur dioxide vs. total sulfur dioxide (r = 0.62)  
• Alcohol koreluje s kvalitou (r = 0.44), density má zápornou korelaci (r = −0.31)

Obsah obrázku text, snímek obrazovky, diagram, Písmo

Obsah vygenerovaný umělou inteligencí může být nesprávný.

Obsah obrázku text, snímek obrazovky, Vykreslený graf, řada/pruh

Obsah vygenerovaný umělou inteligencí může být nesprávný.

**Formální zápis modelu**  
quality = β₀ + β₁·fixed\_acidity + β₂·log\_volatile\_acidity + β₃·log\_citric\_acid + β₄·log\_chlorides + β₅·free\_sulfur\_dioxide + β₆·total\_sulfur\_dioxide + β₇·density + β₈·pH + β₉·sulphates + β₁₀·alcohol + β₁₁·sweet + ε

**Interpretace koeficientů**

* Z modelu vyplývá např.:  
  Alkohol (β ≈ +0.370): Zvýšení obsahu alkoholu o 1 % objemu je spojeno se zvýšením hodnocení kvality vína o 0.37 bodu (p < 0.001).
* Log(volatile acidity) (β ≈ −2.52): Vyšší těkavá kyselost výrazně snižuje hodnocení kvality (p < 0.001).
* Log(chlorides) (β ≈ −1.51): Vyšší obsah chloridů negativně ovlivňuje kvalitu (p < 0.05).
* Sulphates (β ≈ +0.336): Vyšší množství síranů má pozitivní vliv na kvalitu (p < 0.01).
* Sweet (β ≈ +0.231): Sladká vína bývají mírně lépe hodnocena (p < 0.001).
* Některé proměnné (např. density, total sulfur dioxide, pH) nebyly statisticky významné (p > 0.05), i když byly ponechány pro komplexnost modelu.

**Přesnost modelu**

## R² ≈ 0.27: Model vysvětluje přibližně 27 % variability v hodnocení kvality vína.

## RMSE ≈ 0.76: Průměrná odchylka predikce od skutečné hodnoty kvality je přibližně 0.76 bodu.

## Statistická významnost

Pomocí funkce summary(model) bylo zjištěno, že proměnné alcohol, log\_volatile\_acidity, sweet, free sulfur dioxide, sulphates a fixed acidity jsou statisticky významné (p < 0.05). Tyto proměnné mají tedy pravděpodobně skutečný vliv na subjektivní hodnocení kvality vína. Naopak proměnné jako density, pH, log\_citric\_acid a total sulfur dioxide nebyly statisticky významné a lze je případně zvážit pro odstranění při optimalizaci modelu.

# Závěr

Dataset je kvalitní, bez chybějících hodnot. Byly identifikovány odlehlé hodnoty, které byly ponechány. Logaritmická transformace některých proměnných zlepšila jejich rozdělení. Nejvýraznější vliv na kvalitu vína mají obsah alkoholu, těkavá kyselost, sladkost vína, obsah síranů a volného oxidu siřičitého. Lineární model má slušnou predikční schopnost, i když vzhledem k subjektivitě hodnocení nelze očekávat vysokou přesnost. Výsledky odpovídají očekáváním a fyzikálně-chemickým vlastnostem vína.