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 (EDA): 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.

## Transformace dat

Proměnné chlorides, volatile acidity a citric acid měly výrazně nesymetrické (pravostranné) 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)

## Vizualizace

• Histogram kvality ukazuje nejčastější hodnoty 5 a 6.  
• Scatterplot density ~ alcohol ukazuje klesající trend.  
• Korelační matice pomocí corrplot vizualizuje vztahy mezi všemi proměnnými.

# Lineární regrese

## Formální zápis modelu

Model predikující kvalitu vína (quality) byl sestaven jako:  
  
model <- lm(quality ~ alcohol + density + volatile.acidity + chlorides + sulphates + total.sulfur.dioxide + pH, data = vino)

## Interpretace koeficientů

Z modelu vyplývá např.:  
• Alcohol (β ≈ +0.28): Zvýšení obsahu alkoholu o 1 % vol. je spojeno se zvýšením hodnocení kvality o 0.28 bodu.  
• Density (β ≈ −2.5): Vyšší hustota je spojena s nižším hodnocením.  
• Volatile acidity (β ≈ −1.3): Větší těkavá kyselost snižuje kvalitu.

## Přesnost modelu

• R² ≈ 0.28: Model vysvětluje cca 28 % variability hodnocení.  
• RMSE ≈ 0.65: Průměrná chyba predikce je asi 0.65 bodu na škále kvality.

## Statistická významnost

Pomocí summary(model) bylo zjištěno, že proměnné alcohol, density a volatile acidity jsou statisticky významné (p < 0.001). Ostatní proměnné byly na hraně významnosti nebo nevýznamné a lze je 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, hustota a těkavá kyselost. Lineární model má slušnou predikční schopnost, i když vzhledem k subjektivitě hodnocení nelze očekávat vysokou přesnost. Výsledky odpovídají očekáváním a fyzikálně-chemickým vlastnostem vína.