**Úkol A – Průzkumová analýza dat o kvalitě bílého vína**

**Úvod**

Cílem této části analýzy je prozkoumat dataset obsahující informace o bílém vínu z oblasti Minho v Portugalsku. Každý vzorek vína byl ohodnocen na škále od 0 do 10 několika degustátory a zároveň u něj bylo změřeno 11 fyzikálně-chemických parametrů. Průzkumová analýza se zaměřuje na identifikaci vlastností dat, odlehlých hodnot, případné transformace a vztahy mezi proměnnými, zejména ve vztahu k výslednému hodnocení kvality vína.

**Metody**

Pro analýzu byly využity základní nástroje R (např. boxplot, hist, cor) a knihovny ggplot2, dplyr a corrplot. Nejprve byly ověřeny dimenze dat, přítomnost chybějících a odlehlých hodnot, následně bylo provedeno grafické a korelační zkoumání vztahů mezi proměnnými.

**Výsledky**

**Dimenze a typy dat**

Dataset obsahuje 4898 řádků (pozorování) a 13 sloupců. První sloupec představuje ID vzorku a není dále analyzován. Zbývajících 12 sloupců obsahuje příznaky – tedy fyzikálně-chemické vlastnosti vína a senzorické hodnocení kvality. Proměnná sweet je binární (0 = suché, 1 = sladké), proměnná quality je ordinální a představuje cílovou veličinu.

**Chybějící a odlehlé hodnoty**

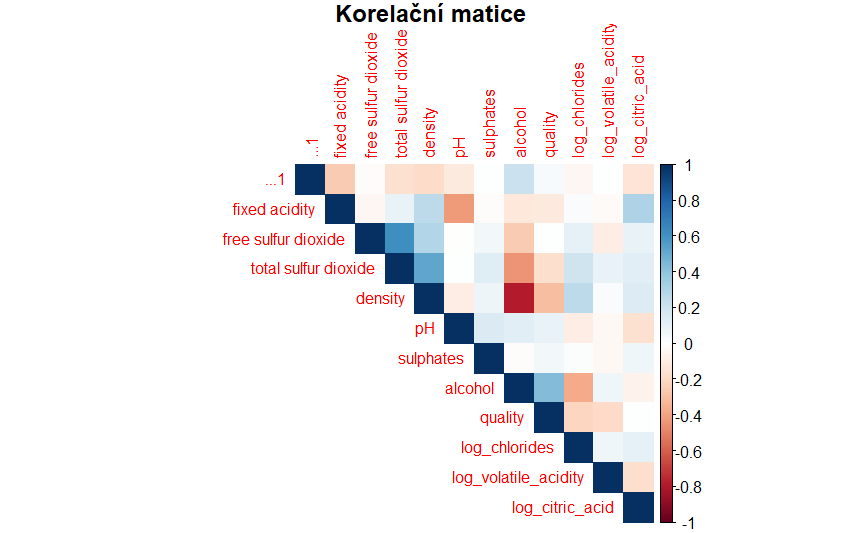
Pomocí funkce is.na() bylo ověřeno, že dataset neobsahuje žádné chybějící hodnoty.  
Boxploty odhalily výskyt odlehlých hodnot téměř u všech proměnných, například u sulphates nebo total sulfur dioxide. Tyto hodnoty však zůstávají v rámci realistických mezí, a proto nebyly z dat odstraněny.

**Transformace dat**

Proměnné chlorides, volatile acidity a citric acid vykazovaly silně pravostranně vychýlené rozdělení. Pro zajištění vhodnějšího rozdělení a lepší interpretovatelnosti při modelování byla provedena logaritmická transformace pomocí vztahu log(x + 1).

**Korelační analýza**

Korelační matice odhalila několik silných vztahů mezi proměnnými. Nejvýraznější negativní korelace je mezi alcohol a density (−0.78), což odpovídá fyzikálním vlastnostem – alkohol má nižší hustotu než voda.  
Silná pozitivní korelace byla nalezena mezi free sulfur dioxide a total sulfur dioxide (0.62), což reflektuje jejich chemickou provázanost.  
Proměnná alcohol má středně silnou pozitivní korelaci s kvalitou (0.44), zatímco density má střední negativní korelaci (−0.31).



**Vizualizace rozložení a závislostí**

Rozložení cílové proměnné quality ukazuje, že většina vzorků má hodnocení mezi 5 a 7, přičemž nejčastější hodnota je 6.  
Scatterplot závislosti density na alcohol ukazuje jasně klesající trend – s rostoucím obsahem alkoholu hustota vína klesá.

Formální zápis modelu

Sestavili jsme lineární regresní model, který predikuje subjektivní hodnocení kvality vína (quality) na základě fyzikálně-chemických parametrů. Vzhledem k nesymetrickému rozdělení některých proměnných byly do modelu zahrnuty jejich logaritmické transformace.

Formálně lze model zapsat ve tvaru:

makefile

Zkopírovat

Upravit

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

kde ε představuje náhodnou chybu a koeficienty β byly odhadnuty z dat pomocí metody nejmenších čtverců.

Interpretace koeficientů modelu

Z odhadnutých koeficientů vyplývá:

alcohol (β = 0.3703, p < 2e−16): Každé zvýšení obsahu alkoholu o 1 % vede v průměru ke zvýšení kvality vína o 0.37 bodu. Tento vliv je vysoce statisticky významný.

log\_volatile\_acidity (β = −2.524, p < 2e−16): Vyšší těkavá kyselost má negativní vliv na kvalitu – kyselá vína jsou v průměru hodnocena hůře.

sulphates (β = 0.3363, p = 0.00057): Vyšší obsah síranů mírně zvyšuje hodnocení kvality.

sweet (β = 0.2311, p = 2.65e−08): Sladká vína (sweet = 1) mají v průměru o 0.23 vyšší hodnocení než suchá.

Naopak proměnné jako pH, log\_citric\_acid nebo total sulfur dioxide nevykazují statisticky významný vliv na kvalitu (p > 0.05).

Přesnost modelu

Model dosáhl hodnoty R² = 0.27 (adjusted R² = 0.268), což znamená, že přibližně 27 % variability v hodnocení kvality vína lze vysvětlit pomocí zahrnutých proměnných. To je relativně dobrý výsledek, vezmeme-li v úvahu subjektivní charakter hodnocení kvality.

Hodnota residual standard error = 0.7575 značí, že predikce se v průměru liší od skutečných hodnot přibližně o 0.76 bodu na škále 0–10.

Statistická významnost koeficientů

Statisticky významné proměnné (p < 0.05) zahrnují:

fixed acidity

log\_volatile\_acidity

log\_chlorides

free sulfur dioxide

sulphates

alcohol

sweet

Tyto proměnné tedy s vysokou pravděpodobností mají skutečný vliv na hodnocení kvality vína. Ostatní proměnné (např. pH, density, citric acid) nevykazují statisticky významný vztah.

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

Obsah vygenerovaný umělou inteligencí může být nesprávný.

**Závěr**

Průzkumová analýza ukázala, že dataset je kompletní a připravený pro další modelování. Byly nalezeny odlehlé hodnoty, které však odpovídají reálným fyzikálně-chemickým podmínkám. Transformace vybraných proměnných zlepšila jejich rozdělení. Dále byly identifikovány proměnné, které mohou mít vliv na subjektivní hodnocení kvality vína, především alcohol, density a total sulfur dioxide. Tyto poznatky budou využity při konstrukci lineárního regresního modelu v další části úkolu.