**Členové:** Jan Hloupý, Růžena Šípková *(seznam členů v týmu)*

**Téma:** Analýza pohádek *(název projektu)*

**Zdroje dat:**

<https://databaze.pohadky.cz/pohadky.csv>

Soubor obsahuje data o jednotlivých pohádkách. Má přibližně 500 řádků. Obsahuje 5 sloupců:

* Id – celé číslo, unikátní identifikátor pohádky (např. 173).
* Název – textový řetězec (např. O červené karkulce).
* Autor – textový řetězec, jméno a příjmení autora, některé pohádky autora nemají (např. Hans Christian Andersen).
* Století – celé číslo, určuje století vzniku, některé pohádky nemají uvedeno (např. 18).
* Země původu – textový řetězec (např. Švédsko).

<https://databaze.pohadky.cz/postavy.csv>

Soubor obsahuje charakteristiku jednotlivých postav. Má přibližně 3000 řádků. Obsahuje 4 sloupce:

* Pohádka – textový řetězec (např. O červené karkulce).
* Postava – textový řetězec (např. vlk).
* Charakter – textový řetězec, určuje kladnost postavy (možnosti: kladná, neutrální, záporná).
* Významnost – textový řetězec, určuje, jako významná postava pro pohádku je (možnosti: málo, středně, velmi)

<https://oblibene-pohadky.cz/hodnoceni>

Online tabulka obsahující celkem cca 5000 řádků stránkovaných po 50. Obsahuje sloupce:

* Datum hodnocení – datum ve formátu dd. mm. yyyy (např. 21. 08. 2017)
* Název pohádky – textový řetězec (např. O červené karkulce).
* Hodnotitel – textový řetězec, určuje označení hodnotitele (hodnotitele podle webu mají unikátní označení, např. PohadkoZrout).
* Hodnocení – celé číslo v rozmezí 1-10, kde 10 je nejlepší hodnocení (např. 8).
* Komentář – textový řetězec, není vyplněný u všech (např. toto je má nejoblíbenější).

<https://www.csfd.cz/zebricky/vlastni-vyber/>

Online tabulka stránkovaná po 100 záznamech. Vyfiltrovaná pouze na kategorii „pohádka“. Lze z ní získat následující údaje:

* Název – textový řetězec (např. Lakomá Barka).
* Země – textový řetězec (např. Československo).
* Režie – textový řetězec (např. Vlasta Pospíšilová).
* Rok – celé číslo (např. 1986)
* Hodnocení – desetinné číslo (např. 92,9).
* Počet hodnocení – celé číslo (např. 2825).

**Charakteristika a cíl:**

Analýza popsaných dat se bude zaměřovat na to jaká je oblíbenost jednotlivých pohádek v závislosti na různých hlediscích. Studie bude určena pro filmové studio, či režiséra, scénáristu za účelem výběru vhodného tématu pro vytvoření nové filmové pohádky. Propojením databáze existujících pohádek s těmi, které byly natočeny a s hodnocením lidí ze dvou různých zdrojů by mohlo přinést zajímavé informace. Úskalím bude správně propojit jednotlivé zdroje.

*(V této části stručně popište problematiku, kterou jste se rozhodli řešit. Vysvětlete, pro koho jsou výsledky vašeho projektu určeny (manažer firmy, trenér týmu, sázkař, starosta města, …) a k čemu by měly přispět?)*

**Klíčové indikátory:**

* Průměrné hodnocení pohádky(pohádek) v závislosti na různých kritériích (země původu, autor, …). Pomůže to vyhledat dobře hodnocené pohádky, které by mohly mít potenciál pro zfilmování. Užitečná bude i možnost filtrovat, zda byla či nebyla zfilmována, případně kolikrát.
* Vytvoření ukazatele, který by reprezentoval vztah mezi hodnocením pohádky a filmu. Tento ukazatel by se pak aplikovat na země či autory a dalo by se například zjistit, u kterých autorů dobré hodnocení pohádky vede i dobrému hodnocení filmu.
* Další ukazatel by se zaměřil na hodnotitele pohádek. Zjistit kteří hodnotitelé dokážou rozpoznat kvalitu pohádky v závislosti na úspěšnosti filmu. Zde by se dalo pak najít dobré hodnocení u nezfilmovaných pohádek.
* Další ukazatel se zaměří na počet hodnocení. Jde o to, že sice může být pohádka hodnocena skvěle, ale neoslovila mnoho hodnotitelů, což pravděpodobně znamená, že nezaujala.
* Dále se analýza zaměří i na postavy. Některé mohou vystupovat napříč více pohádkami. Chceme zjistit, které jsou oblíbené a přinesou pravděpodobnost úspěchu.

*(Zde v alespoň pěti odrážkách uveďte, jaké indikátory byste chtěli zjistit. Zkuste se zamyslet, čím jsou přínosné, proč jsou pro osobu, které je projekt určen, důležité. Dají se jednoznačně interpretovat?)*