Adatelemzés a gyakorlatban Python Pandas és Scikit-learn alapokon

Laczkó Anna

A választott témám az TMDB filmadatbázis adatainak az elemzése. A választásom azért esett erre, mivel régebben sokszor bújtam az IMDB nevezetű filmes adatbázist, nézegettem a statisztikákat, és emiatt úgy érzem, hogy érdekes és izgalmas feladat lesz egy sokkal precízebb, szakmaibb szempontból megvizsgálni. Az adatokat a „Házi_feladatok.pdf” file-ban megadott szempontok közül a következőkkel fogom elemezni:

- Fedezze fel a tévéműsorok népszerűségének trendjeit és fő tényezőit
- Jósolja meg egy tévéműsor sikerét olyan jellemzők alapján, mint a szavazatszám, az átlag és a népszerűség.
- Azonosítsa a legtermékenyebb tévéműsor-alkotókat vagy produkciós cégeket az általuk készített műsorok száma alapján.
- Elemezze a tévéműsorok nyelve és népszerűsége közötti kapcsolatot, és vizsgálja meg a nem angol nyelvű műsorok népszerűségét. Ennél célom, hogy az adatokat tudjam valahogy a nyelvek szerint normalizálni, hogy így mérhető legyen egy angol nyelvű, illetve egy idegen nyelvű sorozat sikeressége egymáshoz.
- Ezek mellett szeretném az adathalmazt alaposabban átvizsgálni és további összefüggéseket keresni.

Az adathalmaz a következő címen található meg: https://www.kaggle.com/datasets/asaniczka/full-tmdb-tv-shows-dataset-2023-150k-shows