**Preciza datelor de filme pentru a înțelege tendințele și preferințele publicului.**

**Cuznetov Mihail, IA-211,III**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Inginerie și Micro-Electronică, Chișinău, Moldova, Informatica aplicată.*

**Absract**

Setul de date IMDb din Top 10.000 de filme acest set de date oferă informații despre filmele apreciate de critici, oferind informații valoroase pasionaților de film, cercetătorilor și cercetătorilor de date. Acesta cuprinde date despre 10.000 de filme clasate printre primele după evaluarea medie a utilizatorilor pe IMDb, o bază de date online populară de filme și emisiuni TV.

Acoperind o gamă largă de genuri, ani de lansare și țări, setul de date oferă o panoramă diversă a istoriei filmului și a cinematografiei globale.

Fiecare film include detalii precum titlul, anul lansării, durata de rulare, evaluarea MPAA, genul(e), regizorul, distribuția principală, evaluarea medie a utilizatorilor, Metascore, numărul de voturi ale utilizatorilor și valoarea brută de box office.

Utile pentru analiză: datele pot fi folosite pentru diverse analize, inclusiv explorarea tendințelor în filmele de top, compararea genurilor și regizorilor și investigarea corelațiilor dintre aprecierea criticilor și succesul de box office.

**Introducere**

Filmele sunt o formă de artă și divertisment care au captivat publicul de secole. De la filme clasice de la Hollywood la producții independente din întreaga lume, filmele ne oferă o perspectivă asupra vieții, a culturii și a lumii înconjurătoare.

În ultimii ani, a existat un interes crescând pentru cercetarea filmelor. Cercetătorii din domeniul științelor sociale, al istoriei și al studiilor culturale au explorat rolul filmelor în societate, precum și impactul lor asupra publicului.

Datasetul IMDb Top 10.000 de filme poate fi utilizat pentru o varietate de aplicații, inclusiv:

Analiza tendințelor în filmele de succes: Acest dataset poate fi utilizat pentru a explora tendințele în genul, regia și actorii filmelor de succes. De exemplu, putem analiza dacă anumite genuri sau regizori au devenit mai populari în timp.

Comparația filmelor din diferite țări: Dataset-ul poate fi utilizat pentru a compara filmele din diferite țări. De exemplu, analiza dacă există diferențe în genurile sau regizorii filmelor populare în diferite țări.

Investigarea relației dintre succesul critic și succesul comercial: Poate fi utilizat pentru a investiga relația dintre rating-ul IMDb al unui film și încasările din box office ale acestuia. Analiza dacă filmele cu un rating IMDb mai mare sunt, de asemenea, mai de succes din punct de vedere comercial.

**Materiale & Metode**

**Setul de date:**

Pentru analiza privind comportamentul clienților în domeniul serviciilor de card de credit, am explorat un set de date detaliat disponibil pe Kaggle, intitulat " IMDB Top 10,000 movies” cu acest set de date care prezintă o listă cuprinzătoare a filmelor de top de la IMDB

Acest set de date oferă o privire extinsă asupra valorilor care definesc succesul unui film și locul său în istoria cinematografiei.

Mai jos se regăsește o prezentare structurată a informațiilor relevante:

**Structura Setului de Date:**

Coloane:

* titlu: numele/titlul filmului.
* year: Anul lansării filmului.
* runtime: durata totală/lungimea filmului în minute.
* certificat: certificarea de vârstă sau evaluarea filmului (de exemplu, PG, R, U/A).
* gen: categoria sau tipul de film (de exemplu, dramă, acțiune, romantism).
* regizor: regizorul filmului.
* vedete: actori și actrițe de prim rang din film.
* evaluare: evaluarea medie IMDB a filmului din 10.
* metascore: evaluarea metascore bazată pe recenzii critice.
* voturi: numărul total de voturi/evaluări ale utilizatorilor filmului primit pe IMDB.
* brut: colectarea totală la box office/veniturile brute ale filmului.

Sursă:

Acest set de date a fost organizat prin tehnici de scraping web, extragând date direct de pe site-ul IMDB. Deși IMDB rămâne una dintre cele mai de încredere surse de date legate de filme, este esențial să ne amintim că web scraping poate implica discrepanțe, iar datele ar trebui utilizate în mod responsabil.

**Standardizarea datelor**

Scanând datele, există câteva probleme care trebuie rezolvate înainte de analiză:

1. Cadrul de date conține celule goale în mai multe coloane
2. Coloana „an” are formate diferite pentru ani
3. Coloanele numerice au tip șir de date în loc de tipuri de date numerice
4. Coloanele „director” și „stele” conțin liste de nume sub forma [„A”, „B”, „C”, ...] unde A, B și C sunt nume, dar nu au un listează tipul de date <br>

Pentru a remedia aceste probleme:

1. Toate celulele goale vor fi umplute cu valoarea implicită lipsă NaN a lui Pandas
2. Folosind lista de înțelegere, coloana „an” poate fi standardizată în format AAAA
3. Convertiți coloanele numerice în tipuri de date numerice
4. Convertiți coloanele „director” și „stele” într-un tip de date de listă

**Gestionarea valorilor lipsă**

Deși celulele goale au fost umplute cu valori NaN, valorile lipsă trebuie încă tratate înainte de analiză.

* title 0.000000
* year 0.000000
* runtime 0.000000
* certificate 3.614580
* genre 0.000000
* director 0.000000
* stars 0.000000
* rating 0.000000
* metascore 20.032491
* votes 0.000000
* gross 28.774495
* dtype: float64

Privind rezultatul de mai sus, numai coloanele „metascore” și „gross” au o proporție semnificativă de valori lipsă.

Pentru a gestiona aceste valori lipsă, se va compara un model de regresie liniară și K-Nearest Neighbors, iar valorile lipsă vor fi imputate folosind modelul mai performant.

**Analiza exploratorie a datelor**

* title object
* year int64
* runtime int64
* certificate object
* genre object
* director object
* stars object
* rating float64
* metascore float64
* votes int64
* gross float64
* pred\_metascore float64
* pred\_gross float64
* dtype: object

**Codul disponibil**

Întregul cod pentru executarea operațiilor,și pentru crearea modelului de regresie logistcă, este dispobil pe link-ul: <https://github.com/NikMineZx/Analiza-Datelor>

**Rezultate:**

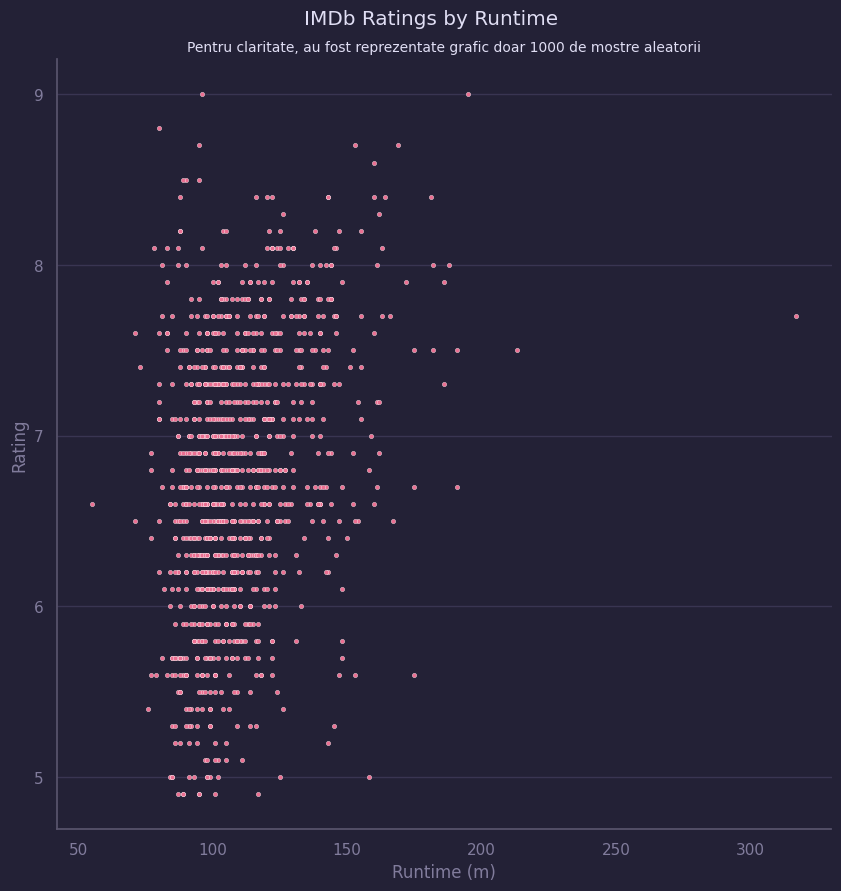
****

Figure 1 IMDb Reitingul pe Durata

Privind coeficientul de corelație dintre „durată de rulare” și „evaluare” și tendința generală a intrigii de mai sus, se pare că oamenii tind să se bucure de filme mai lungi. Cu toate acestea, această relație este slabă. Mai mult, ar putea exista variabile confuze care explică această relație. De exemplu, filmele mai lungi pot avea tendința de a avea investiții mai bune și o calitate mai bună a producției și, prin urmare, sunt evaluate mai bine.

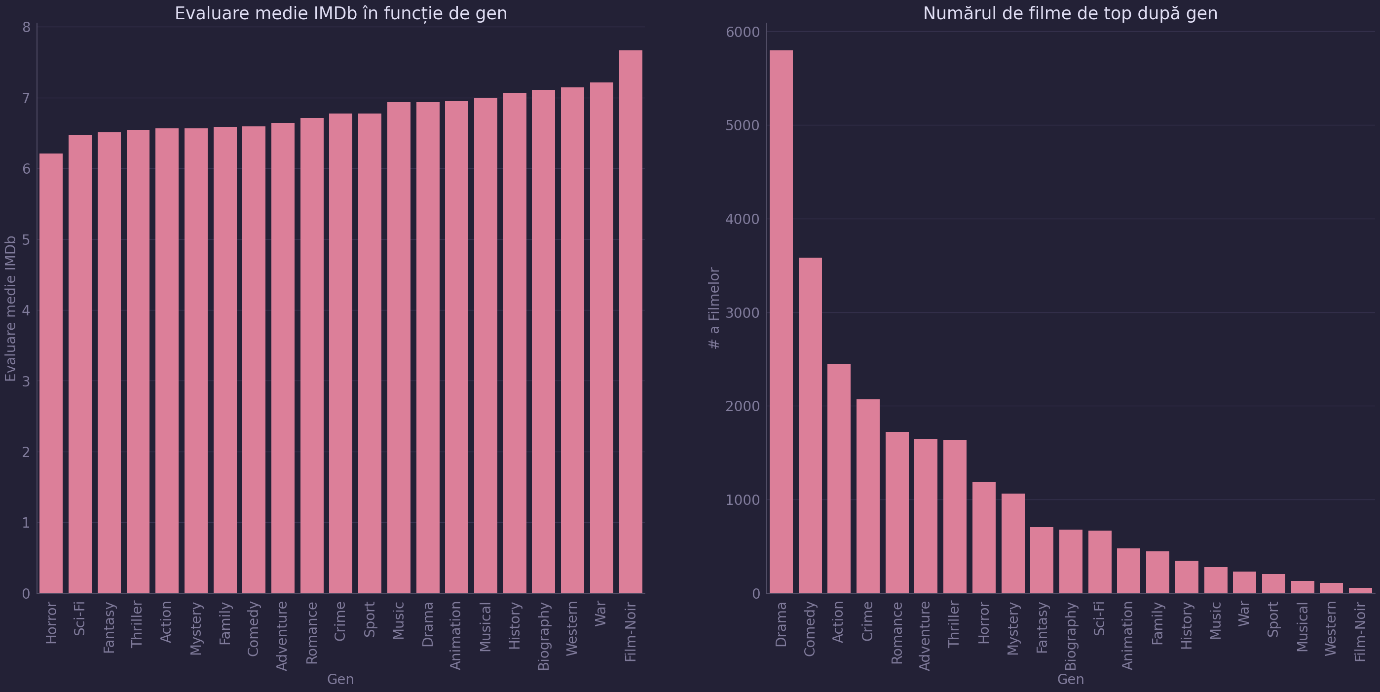


Figure 2 Evaluarea medi IMDB in functie de gen / Numarul de filme de top dupa gen

Se pare că mai multe genuri de nișă, cum ar fi film-noir, western și istorie, tind să fie evaluate mai sus decât genurile mai populare precum horror, fantezie și acțiune. Acest lucru s-ar putea datora modului în care aceste filme sunt comercializate; Genurile de nișă pot fi comercializate numai către piețele care au o apreciere cunoscută pentru gen, în timp ce genurile populare sunt comercializate către populația generală. Astfel, genurile de nișă nu vor avea atât de mulți oameni să își evalueze prost filmele.

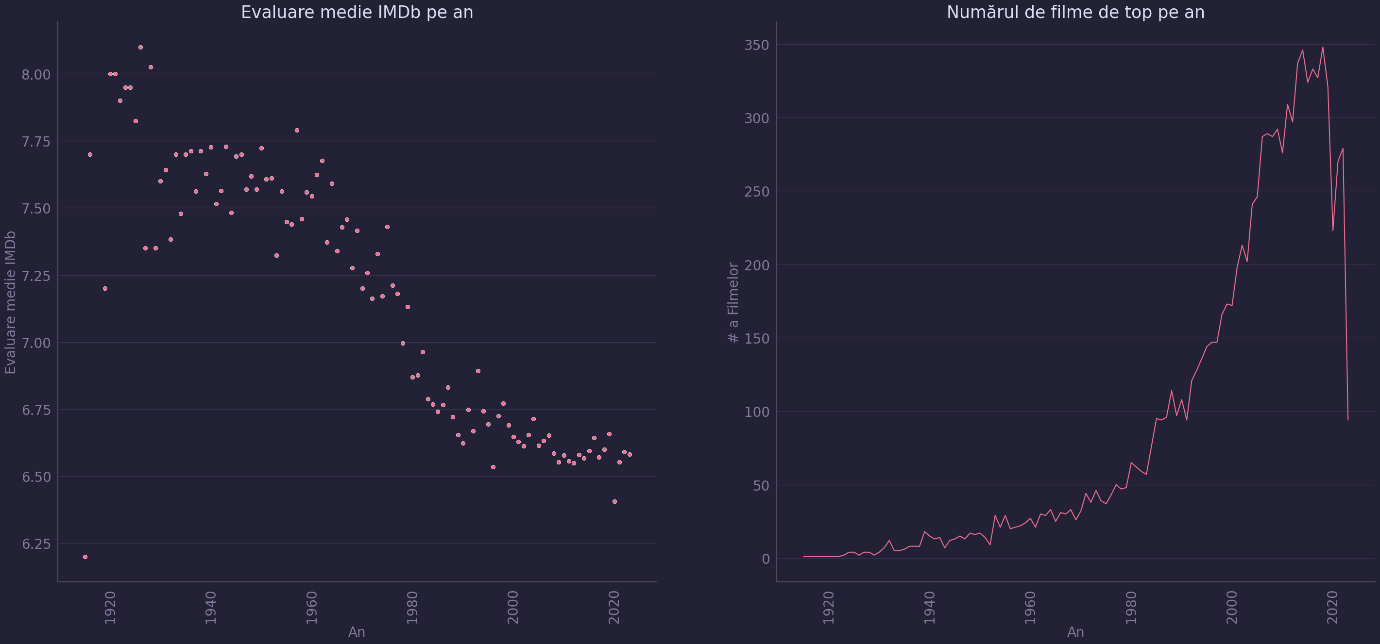


Figure 3 Evaluare medie IMDb pe an / Numarul de filme de top pe an

Deși numărul de filme din primele 10.000 a crescut constant de-a lungul anilor, ratingul mediu al acestor filme a scăzut constant de-a lungul anilor. În general, filmele mai vechi sunt evaluate mai bine decât filmele mai noi.

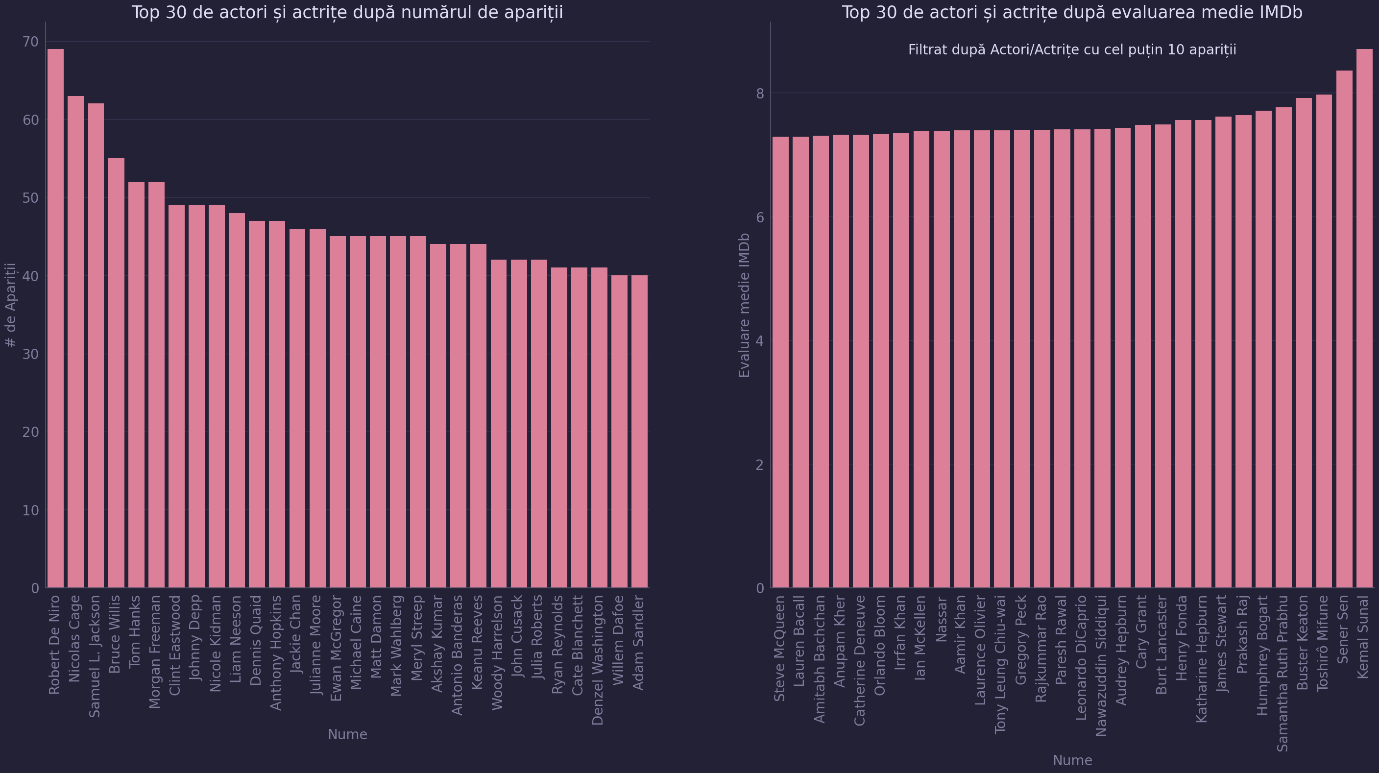


Figure 4 Top 30 de actori si actrite dupa numarul de aparitii/medie IMDb

Comparând cele două intrigi, se pare că actorii / actrițele europeni și americani tind să apară în cele mai multe filme de top, în timp ce actorii / actrițele din Asia de Est și India tind să apară în filme cu cote mai bune. Diferența de număr de apariții s-ar putea datora faptului că filmele de la Hollywood, care joacă predominant europeni și americani, tind să aibă mai multe investiții financiare decât omologii lor din Est și astfel, sunt lansate filme de la Hollywood mai bine produse.

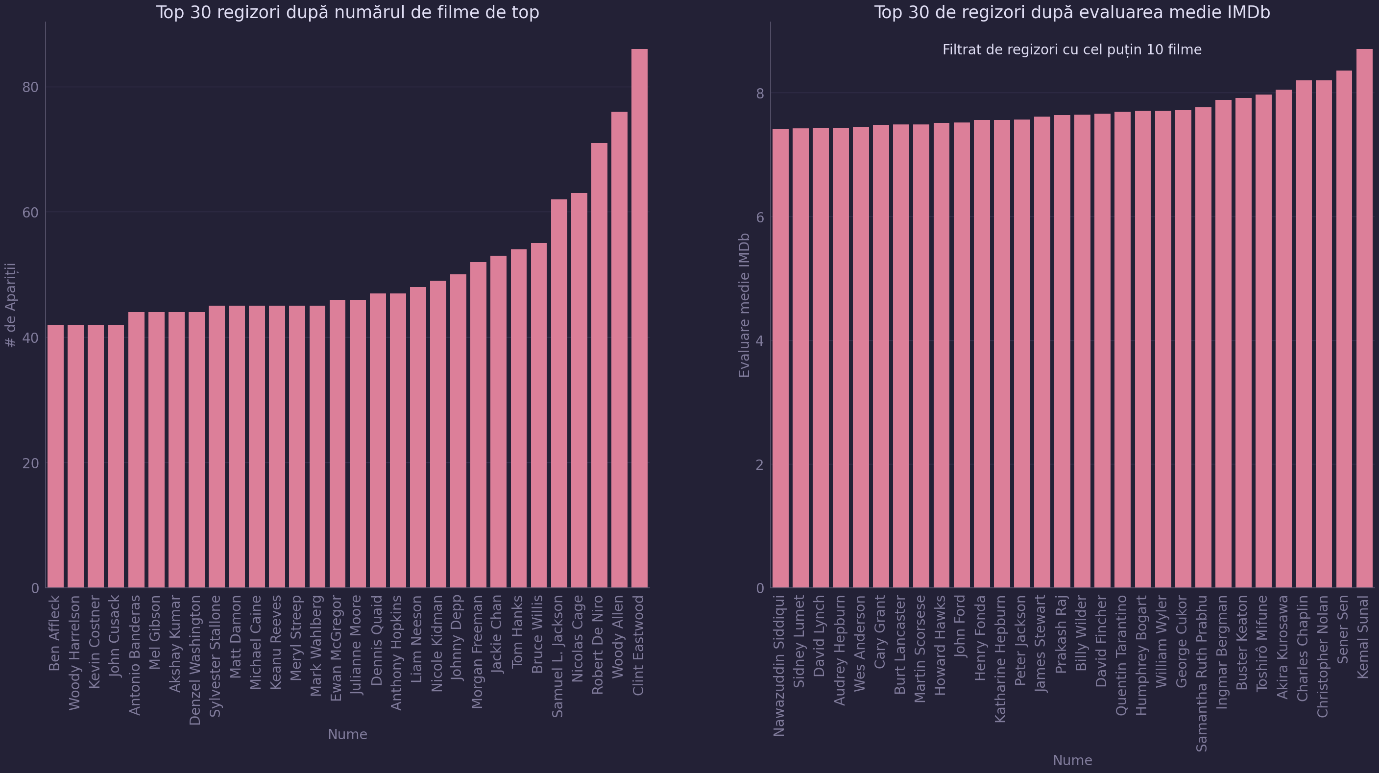


Figure 5 Top 30 regizori dupa numarul de filme de top/ medie IMDb

Privind la intriga din stânga, regizorii occidentali precum Clint Eastwood și Robert De Niro tind să fi regizat majoritatea filmelor de top pe IMDb, în timp ce intriga din dreapta arată că regizorii estici precum Kemal Sunal și Akira Kurosawa fac cele mai bine cotate filme. Intriga din stânga arată, de asemenea, o predispoziție față de regizorii care au acționat și ei, în timp ce intriga din dreapta nu pare să prezinte aceeași predispoziție.

|  |  |  |  |
| --- | --- | --- | --- |
|  | MSE | R^2 |  |
| Regresia liniară Metascore | 149.70511373420692 | 0.5007831093926833 | Modelul de regresie KNN are performanțe puțin mai bune cu o eroare pătratică medie mai mică și o valoare R^2 mai mare. Deci, modelul de regresie KNN va fi utilizat pentru a imputa valorile Metascore. |
| Metascore K-Cei mai apropiați vecini | 150.47799022573147 | 0.4982058226901902 |
| Regresia liniară a veniturilor brute | 3023.9001045960417 | 0.30846952959416596 | Din nou, modelul de regresie KNN depășește modelul de regresie liniară cu un MSE mai mic și R^2 mai mare. Deci, modelul de regresie KNN va fi folosit și pentru a imputa valoarea câștigului brut. |
| Câștiguri brute K-Cei mai apropiați vecini | 2658.022211867578 | 0.3921415103203314 |

Accuracy: 0.805574912891986

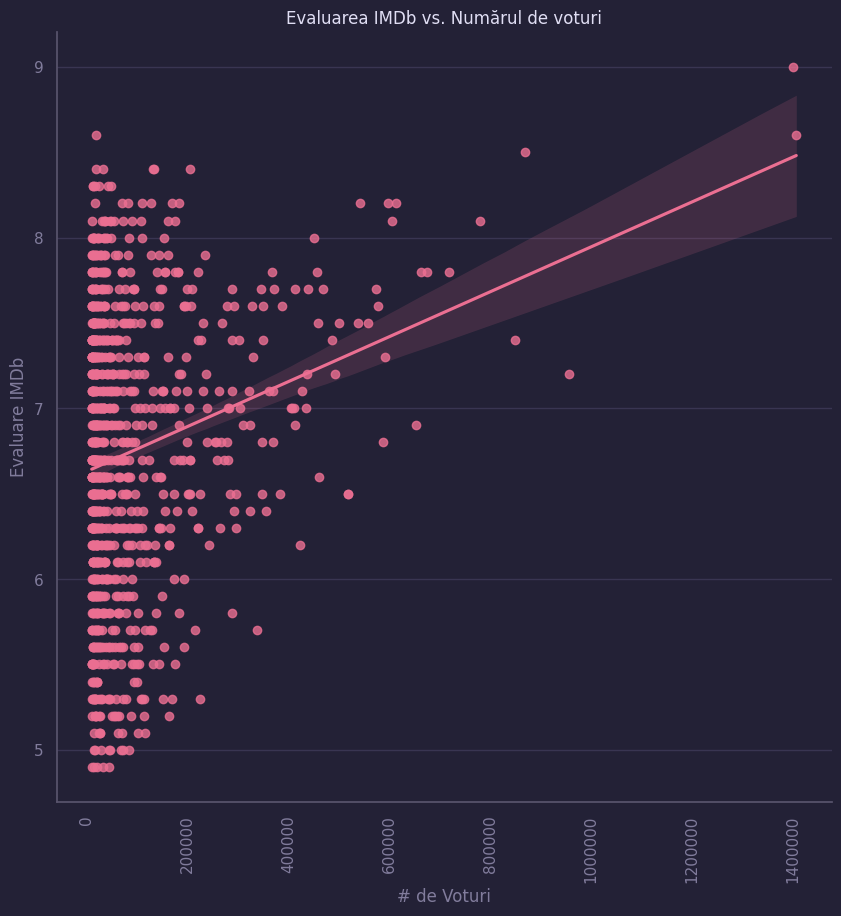


Figure 6 Evaluarea IMDb vs. Numarul de voturi

Filmele care obțin mai multe voturi tind să aibă și un rating IMDb mai ridicat. Deoarece numărul de voturi este probabil proporțional cu popularitatea unui film, este sigur să spunem că filmele mai populare tind să fie evaluate mai sus decât filmele mai puțin populare.

**Bibliografie**

1. Librarie: Pandas [https://pandas.pydata.org\](https://pandas.pydata.org\\)
2. Librarie: matplotlib https://matplotlib.org
3. Librarie: sklearn.model\_selection https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html
4. Librarie: Grafica matlplotlib https://github.com/h4pZ/rose-pine-matplotlib
5. Librarie: Regresia liniara python https://www.kaggle.com/code/mosnoiion/regresia-liniara-popularitatea-cantecelor
6. Librarie: Data setul: https://www.kaggle.com/datasets/ashutoshdevpura/imdb-top-10000-movies-updated-august-2023