# Čo robí filmy (ne)obľúbenými?

sVedkovia Dátovovi

8.1.2024

Andrej Čerňanský
Terézia Füleová
Terézia Koleková
Mária Kopecká
Dávid Pitoňák

# Obsah

# Úvod

Každý človek má nejaké obľúbené alebo naopak neobľúbené filmy. Ak sa subjektívne hodnotenie obľúbenosti filmu viacerých ľudí spojí, vzniká akési objektívne hodnotenie. Okrem takéhoto hodnotenia (ratingu) má každý film rozličné iné parametre. Táto práca sa zameriava na analýzu faktorov, ktoré môžu ovplyvniť popularitu filmov.

Preskúmame rôzne aspekty, ktoré môžu vplývať na obľúbenosť alebo aj neobľúbenosť filmov. Tieto faktory zahŕňajú napríklad žáner filmu, štúdio, ktoré ho produkovalo, časové obdobie uvedenia filmu na trh, a samozrejme, zúčastnených hercov a členov produkčného tímu. Okrem toho budeme skúmať aj iné možné faktory, ktoré by mohli ovplyvniť úspech filmu, ako je rozpočet filmu, dĺžka jeho trvania alebo rodičovské pomôcky ako level prítomnosti nahoty, násilia, alkoholu, zastrašovania alebo nadávok vo filme..

V záverečnej časti tohto projektu sa pokúsime identifikovať a zhodnotiť hlavné faktory, ktoré prispievajú k úspechu alebo neúspechu film. Táto naša analýza môže slúžiť ako malý návod pre vyprodukovanie úspešného a obľúbeného filmu.

# 1. Špecifikácia otázok

V tejto časti sa budeme venovať špecifikácii hlavných otázok, ktoré nás viedli k realizácii tohto projektu. Cieľom je jasne a presne definovať otázky, ktoré nás zaujímajú, aby sme mohli systematicky analyzovať faktory ovplyvňujúce popularitu filmov a ich príčiny.

Hlavné otázky, ktoré budeme skúmať, zahŕňajú:

- Ako žáner filmu ovplyvňuje jeho obľúbenosť alebo naopak neobľúbenosť?
- Má štúdio a producenti vplyv na úspech filmu?
- Ako časové obdobie uvedenia filmu na trh ovplyvňuje jeho prijatie? Môžu historické kontexty ovplyvniť úspech filmu?
- Môže prítomnosť členov produkčného tímu prispieť k obľúbenosti filmu?
- Aký vplyv má rozpočet alebo dĺžka trvania filmu na jeho popularitu?
- Ovplyvňuje množstvo násilia, zastrašovania, sexuality, nadávok alebo alkoholu obľúbenosť filmu?

Tiež sa pokúsime zoradiť tieto faktory podľa vplyvu a zistíme tak najviac a najmenej dôležitý faktor obľúbenosti filmu.

# 2. Prehľad použitých dát

V tejto časti sa budeme venovať detailnému prehľadu dát, ktoré sme použili na realizáciu tohto projektu. V projekte sa obmedzujeme na informácie získané zo stránky IMDB, ktorá predstavuje významný zdroj dát v oblasti filmového priemyslu. IMDB nám ponúka bohatú zbierku informácií o filmoch, vrátane názvu filmu, celkovej dĺžky filmu, roku vydania, priemerného hodnotenia, počtu hodnotení, druh žánrov, mená režisérov, mená scénaristov , mená hercov, názvy krajín pôvodu, názvy produkčných firiem, odhadovaného rozpočtu, celkového hrubého zisku a ďalších... Na podstránke /parentalguide sa nachádza aj pomôcka pre rodičov typu koľko sexuálnych a násilných scén sa vo filme nachádza.

Niektoré z informácií o filmoch sú voľne dostupné na stiahnutie pre nekomerčné účely avšak viacero zaujímavých dát ponúkajú len cez platené api. Preto sme sa rozhodli stránku zoscrapovať. Pri scrapovaní sme museli použiť knižnicu Selenium a ChromeDriver aby sme mohli parsovať príslušný html kód. Tento proces si opíšeme viac v časti 4.Scrapovanie stránky imdb.com.

Okrem scrapovania, sme využili aj spomínané voľne dostupné dáta, ktoré boli stiahnuté zo stránky https://developer.imdb.com/non-commercial-datasets/. Ich kópiu sme si lokálne stiahli do počítača, keďže dáta sú na stránke pravidelne updatované. Kvôli obrovskej veľkosti jednotlivých datasetov, sme použitím Jupyter Notebooku vyfiltrovali dáta jednotlivých datasetov, aby sme ich mohli uploadnúť na GitHub, z ktorého ich stiahneme do Google Colabu, kde s nimi budeme pracovať.

Využité datasety:

- title_basics.tsv - v tomto datasete sa nachádzajú základné dáta o seriáloch a filmoch (názov, dĺžka filmu, rok vydania, žáner a iné atribúty); vyfiltorvali sme tieto dáta iba na filmy a vyhodili nepotrebné stĺpce
- title_ratings.tsv - tento dataset obsahuje priemerné hodnotenie filmu a takisto počet hodnotení pre daný film; filmy na seba odkazujú pomocou tconst stĺpca
- title_crew.tsv - tieto dáta pridávajú informáciu ku každému filmu, kto ho režíroval a kto ho napísal vo forme id; pomerne veľa hodnôt v tomto datasete chýba, no stále sa dá použiť na analýzu
- name_basics.tsv - dataset, vďaka ktorému vieme presnejšie identifikovať režisérov a scénaristov; obsahuje meno, rok narodenia a úmrtia, profesia/e a zopár filmov, v ktorých daný človek pôsobil

Datasety title_basics.tsv, title_ratings.tsv a title_crew.tsv boli pospájané do jedného datasetu s názvom movies_v2.tsv, a to pomocou jednotlivých identifiátorov pre filmy (tconst). Keďže v title_basics.tsv sme vyfiltrovali dáta iba na filmy, takto sme vedeli napojiť ostatné dva datasety na tento, s tým, že sa zachovali iba relevantné dáta pre naše filmy a ostatné sa zanechali. To výrazne zmenšilo objem datasetu.

Pre dataset name_basics.tsv boli vyfiltrovaný ľudia, ktorý sa vyskytli niekde medzi režisérmi alebo scénaristmi na už vyfiltrovaných dátach o filmoch. Tento dataset sme pomenovali dirs_writrs.tsv.

Proces filtrovania, je priložený ako príloha k nášmu projektu s názvom extraction.ipynb.

## Exploratívna analýza voľne dostupných dátach

Keď si prezrieme všetky stĺpce teda: 'tconst', 'primaryTitle', 'originalTitle', 'isAdult', 'releaseYear', 'runtimeMinutes', 'genres', 'averageRating', 'numVotes', 'directors' a 'writers'. 'Tconst' je vždy unikátny identifikátor. 'primaryTitle' a 'originalTitle' sú vo väčšine prípadov rovnaké, ale vyše 60000 je rôznych, čo je pomerne veľké číslo. Vo väčšine stĺpcov nechýba žiadna hodnota okrem: runtimeMinutes, writers a directors. Medián počtu hlasov je 67. Väčšina filmov nie je len pre dospelých, najčastejší rok je 2019, čas trvania filmu 90 minút, žánrom je dráma a najviac hodnotitelia pridávali 6.2 hviezdičky, ale zato bolo najčastejšie len 9 hlasov. Ďalej sme zoscrejpovali hercov, krajinu pôvodu, použitý buget, a rodičovské pomocky a potom sme pracovali aj s nimi.

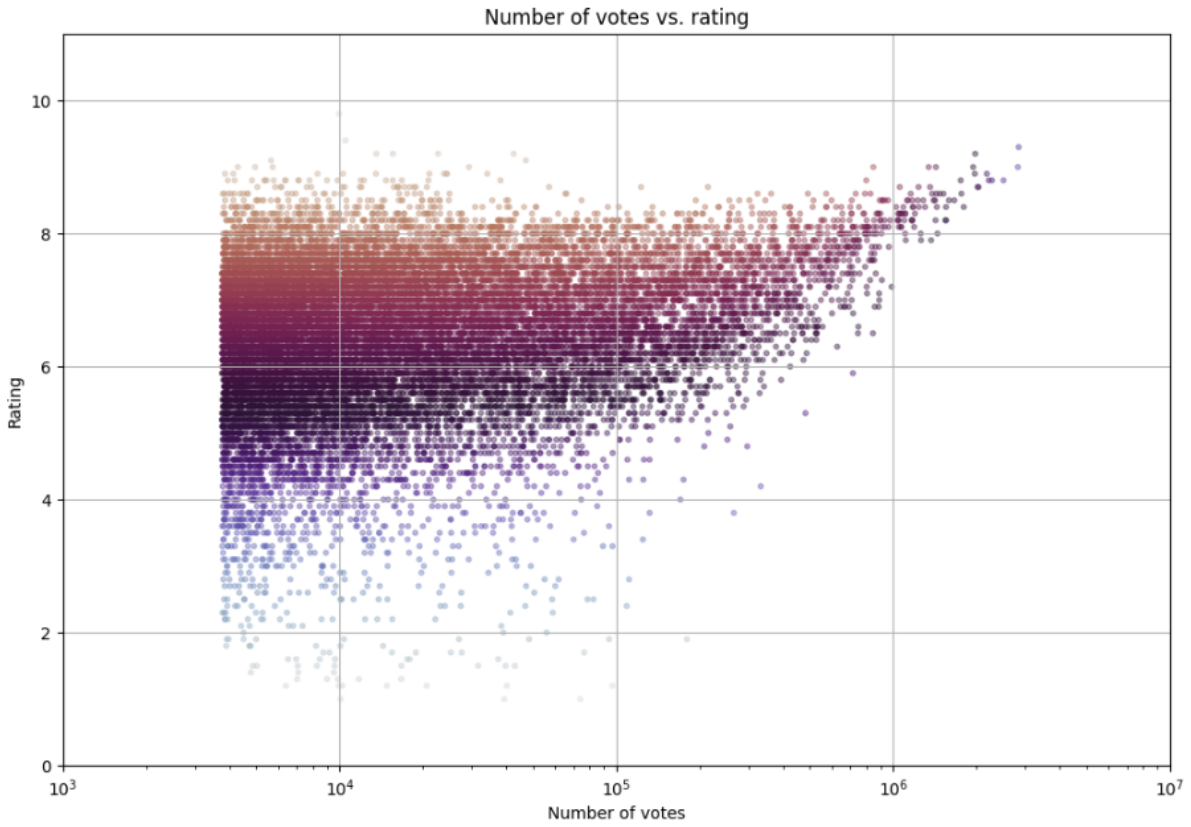
Skúsili sme tiež vypočítať pearsonovu a spearmanovu koreláciu medzi číselnými atribútmi. Pearsonova nám moc nepovedala alebo teda jej hodnoty boli moc blízke k nule. Keď sme vyskúšali spearmanovu tak sa objavila mierna korelácia medzi časom trvania filmu a počtom hlasov. Ak sme si filmy obmedzili len na tie, čo majú počet hlasov väčší ako je priemer, tak nám vyšla vyplynul možný súvis medzi počtom minút vs hodnotenie a rokom vzniku vs hodnotenie, čo budeme aj ďalej skúmať.

# 3. Postup a výsledky analýzy

## 3.1. Analýza na voľne dostupných dátach

### Vplyv počtu hlasov na rating

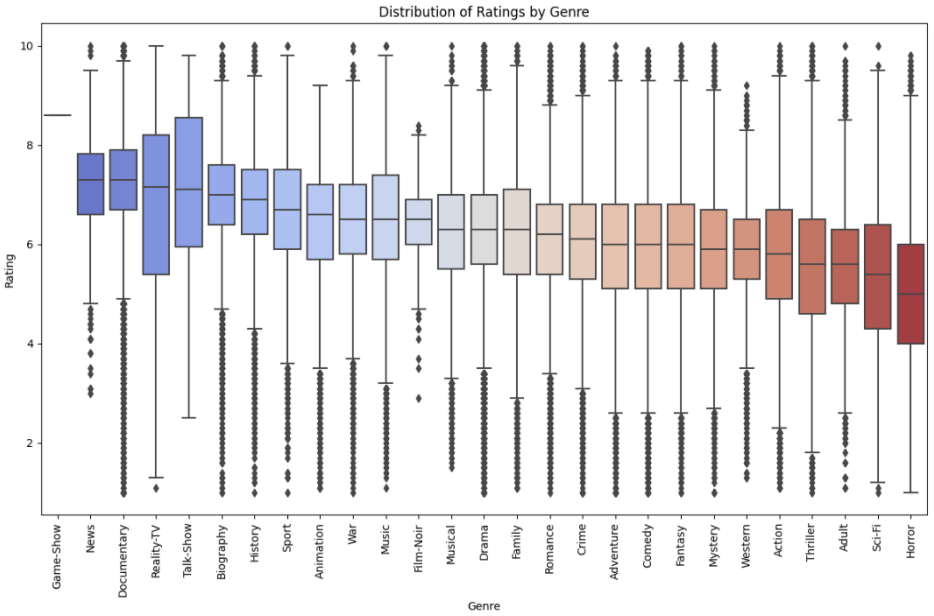
Naša dedukcia je, že výrazne obľúbené alebo výrazne neobľúbené filmy majú aj výrazne väčší počet hlasov, lebo oslovili sledovateľa alebo hodnotiteľa do takej pozitívnej alebo negatívnej miery až sa neunúval, vyhľadal film na imdb a náležite ho ohodnotil. Z korelačnej mapy vyššie by mal vyhrávať o čosi pozitívny nad negatívnym dojmom z filmu. Je milé, že v spoločnosti sa nájde stále viac ľudí čo idú filmy chváliť ako zhadzovať. Nasledujúci graf ukazuje teda vzťah medzi počtom hlasov a ratingom. Obmedzili sme sa na filmy, ktoré majú hlasov viac ako jeho priemer v dátach. Zvolili sme logaritmickú škála pre os x (počet hlasov), lebo rozsah hodnôt počtu hlasov je veľmi široký. Logaritmická transformácia nám umožňuje lepšie vizualizovať rozdiely medzi filmami s rôznymi počtami hlasov bez toho, aby extrémne hodnoty dominovali v grafe. Farebne je zobrazená lineárna regresia na týchto dátach. Čím je bod v grafe tmavší, tým je bližšie ku krivke lineárnej regresie.
Vidíme, že filmy s nadpriemerným počtom hlasov majú spravidla vyššie ratingy.



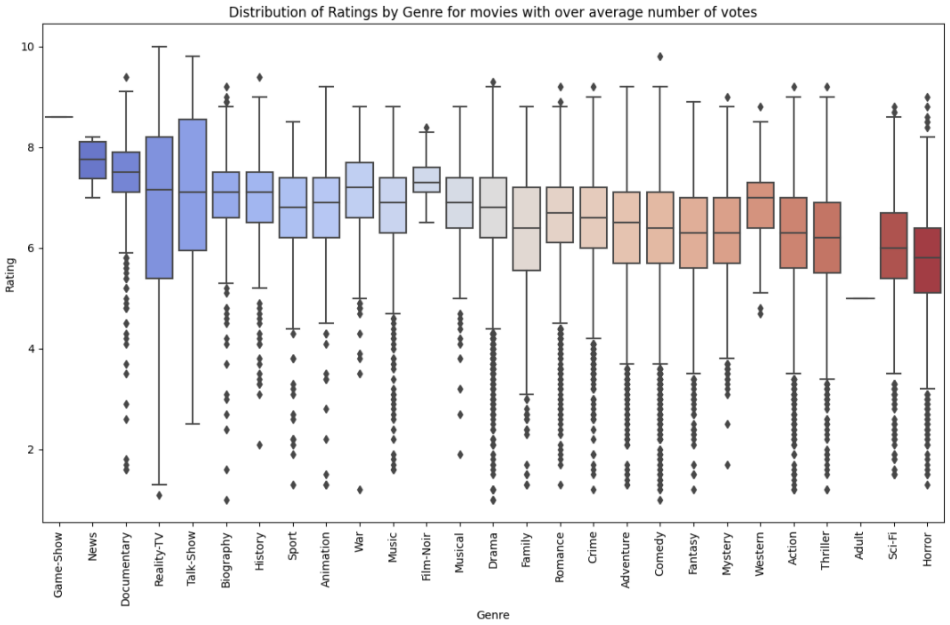

Tento fakt považujeme ako prvý smerodajný pri určení kritéria na obľúbenosť filmu.

## Vplyv žánrov filmov

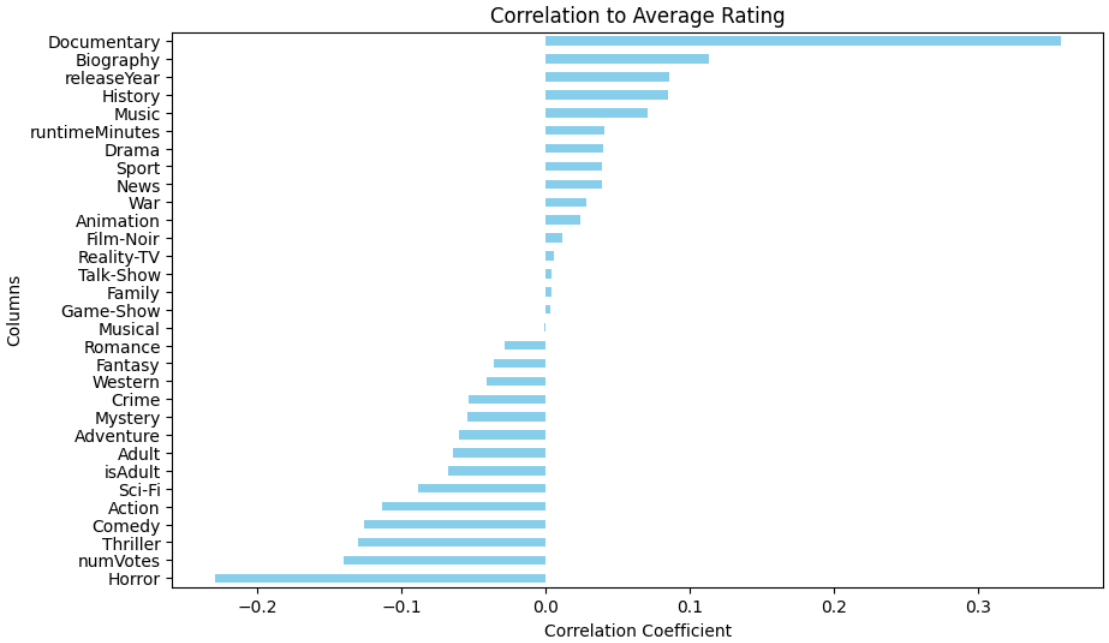
V tejto časti sme si pre každý žáner urobili boxplot distribúcie ratingu daného žánru. Prvý graf je na všetkých dátach z movies zoradený podľa strednej hodnoty. Je zaujímavé pozorovať jednotlivé veľkosti distribúcií. Na prvý pohľad nás zaujal žáner 'Documentary', ktorý má pomerne malú distribúciu a vysoký medián. Tiež je zaujímavý žáner, ktorý má medián najnižší - 'Horror', lebo jeho dolná hranica Q1 siaha takmer k nule. Spolu so žánrom 'Sci-fi' ich predbežne nominujeme za najneobľúbenešie žánre.




V tomto druhom grafe, sme zachovali poradie ako v prvom, ale zobrali sme len filmy s nadpriemerným počtom hlasov. Zaujímavé zmeny pozorujeme pri westerne, kde výrazne stúplo mediánové hodnotenie. Tiež žáner 'News', kde sa distribučný interval dosť zmenšil. Toto však pripisujeme okliešteniu dát. Oproti predošlému grafu má žáner 'Film Noir' vyššie mediánové hodnotenie.



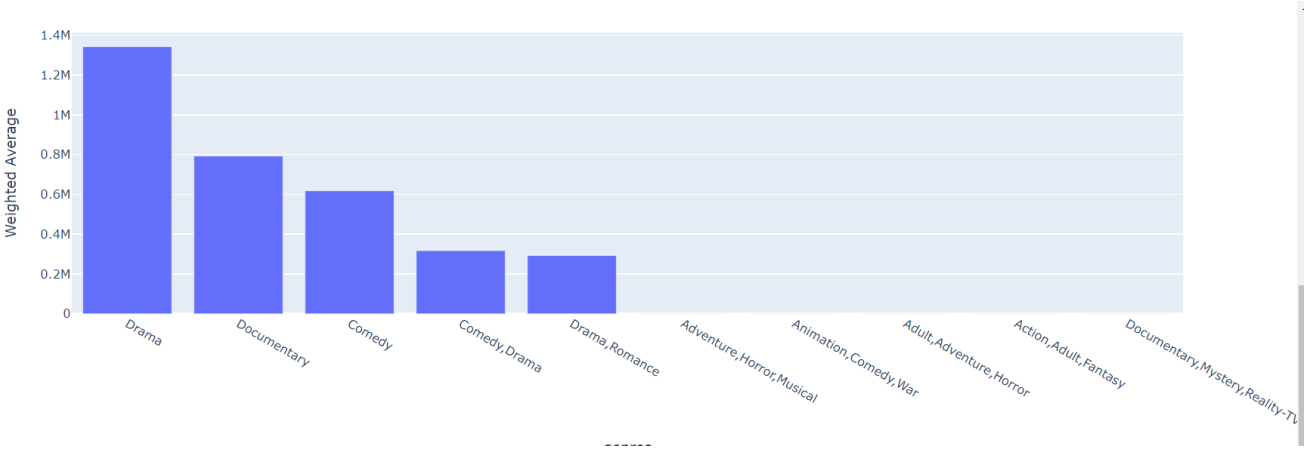
Vykonali sme pozorovanie korelačných koeficientov jednotlivých číselných premenných s ratingom, pomocou Spearmanovho korelačného koeficientu, nakoľko nemáme záruku normality. Jednotlivé žánre boli upravené na numerické premenné zavedením tzv. Dummy variables). Pri žánroch Documentary 0.35, Biography 0.11 možno badať pozitívny rast.. Najzásadnejšie poklesy ratingu boli pozorované pri: horrore -0.23, počet hlasov -0.14.




Štatistickým testovaním sme sa snažili prísť na žanre, ktorých filmy majú významne vyššiu distribúciu ratingov filmov, teda priemerne vyššie ratingy. Konkrétne sme použili test Mann-Whitney U test na test rovnosti dvoch distribúcií, vždy sme porovnávali ratingy pre jeden konkrétny žáner ku všetkých ostatným filmom, len pre filmy s nadpriemerným počtom hlasov. Nulová hypotéza bola, že distribúcia konkrétneho žánru sa od ostatných žánrov nelíši. Alternatívna hypotéza bola, že distribúcia ratingov konkrétneho žánru je významne vyššia (alternative 'greater'). Pre žánre, ktoré vyšla p-hodnota menšia ako 5% môžeme odhadovať, že pravdepodobne sú obľúbenejšie, naopak vysoká p-hodnota (blízko 1) vraví, že sme mali použiť alternatívu 'less', lebo sa jedná o neobľúbené, resp. nekvalitné žánre. Medzi obľúbené žánre môžeme teda považovať: Drama, Documentary, War, Biography, History, tie mali najnižšiu p-hodnotu. Naopak podpriemerné žánre sú Sci-fi, Horror, Fantasy, Comedy, Mystery, Thriller, Action.

## Ako pôsobí kombinácia žánrov?

V našej databáze máme 1275 druhov kombinácií žánrov. Celkovo chýba asi 10090 hodnôt, ale v porovnaní s veľkosťou databázy je to zanedbateľné množstvo. Žánre sme ohodnotili podla pomeru hlasov a hodnotenia a usporiadali. Vyšli nám nasledujúce hodnoty:

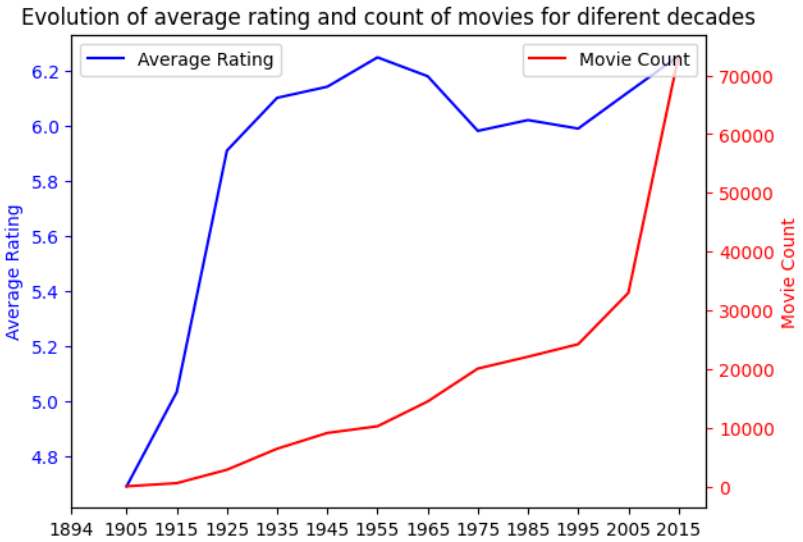



z čoho vyplýva, že príliš zvláštne kombinácie žánrov majú nižšie priemerné hodnotenie ako tie viac bežné.

## Vplyv roku vydania filmov

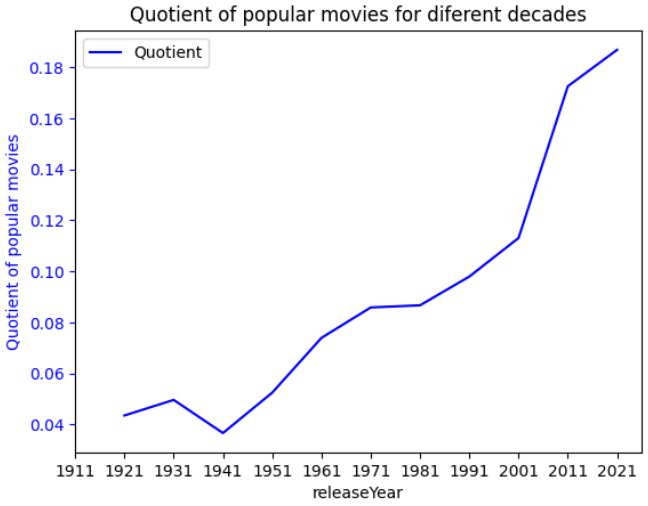
V tejto časti sme robili sumárne štatistiky pre filmy z jednej dekády pomocou metódy pohyblivého okna. V grafe bodu zodpovedá vždy prechádzajúca dekáda.
Pri tejto analýze sme brali do úvahy všetky filmy, nakoľko je zrejmé, že hlavne pre tie z dávnej minulosti nebude toľko hlasov.

Na nižšie uvedenom grafe môžeme vidieť jednak vývoj počtu filmov za jednotlivú dekádu, čo hlavne v posledných desaťročiach prudko stúpa. Taktiež vidíme, že medzi obdobia z najvyšším priemerným ratingom medzi filmami patria 1945-1955, teda obdobie po 2 sv. vojne a 2005-2015.




Nižšie uvedený graf popisuje podiel filmov s ratingom nad 7.5 ku všetkým filmom vzniknuvších v danom období. Najlepšie je opäť súčasné obdobie 2011-2021. Vidíme, že v

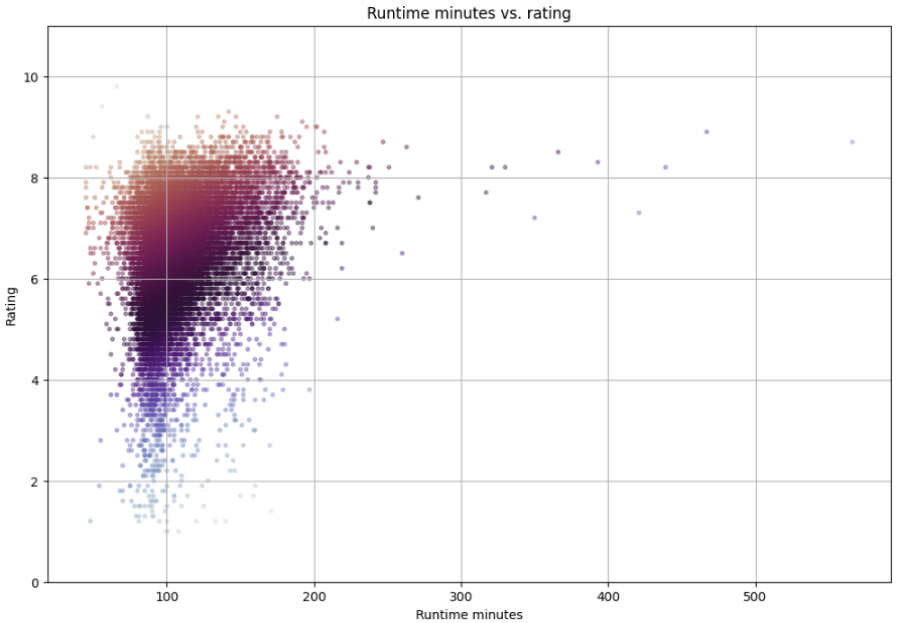
medzivojnovom období je len málo takých filmov. Ale od konca druhej svetovej vojny môžeme sledovať takmer stály rast.



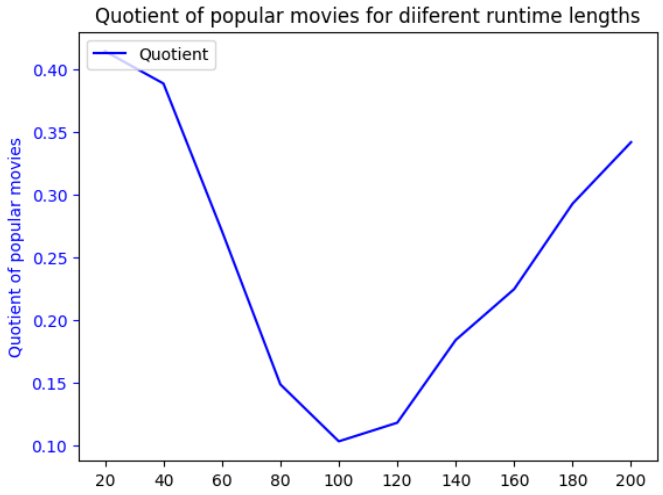

## Vplyv dĺžky filmu

Neočakávali sme, že pri prvotnom skúmaní dát nám pri Spearsonovej korelácií vyjde (aj keď malá ale predsa) závislosť ratingu aj od dĺžky filmu. V tejto časti sa pozrieme na túto závislosť trochu bližšie. Nasledujúci graf ukazuje vzťah ratingu a dĺžky pre filmy, ktoré mali počet hlasov väčší ako priemer. Pre túto časť sme odstránili filmy s dĺžkou nad 1000 minút. Farebne je znázornená lineárna regresia, teda najtmavšie body zodpovedajú filmom, ktoré sú krivke lineárnej regresie najbližšie.

Môžeme si napríklad všimnúť, že filmy trvajúce viac ako 2 hodiny majú málokedy nízke hodnotenie v porovnaní s celkom. Filmy s kratšou dĺžkou sa sústreďujú trošku viac do ľavého horného rohu, a teda naše prvé tvrdenie z tohto grafu je, že obľúbenejší je skôr dlhý alebo krátky film. Tie so strednou dĺžkou môžu byť aj obľúbené aj neobľúbené.



Čo sa týka dĺžky a pomeru filmov v danej dĺžke s ratingom nad 7.5 ku všetkým, vychádzajú najlepšie krátke filmy (20-40 min), ktorý takmer 40 percent má rating nad 7.5. Pre klasické celovečerné filmy okolo dvoch hodín to je najhoršie, okolo 15 percent, ale následne od dĺžky dvoch hodín to lineárne rastie.




## Analýza producentov a režisérov

Tu sa pozrieme na celkový prehľad tvorcov filmov. Dáta sú miestami nevyplnené a teda informácia o režiséroch nie je pri 3564 filmoch a informácia, kto vytvoril scenár chýba pri 37422 filmoch. Pre naše dáta, sme zistili, že na všetkých filmoch sa podieľalo 340 777 režisérov, z toho 134 800 je unikátnych. Pri tvorení scenára filmov sa podujalo 509 554 scenáristov, z toho 210 724 je unikátnych.

V nasledujúcej tabuľke vidíme prvých 8 režisérov, ktorí mali účasť na najviac filmoch na poste režisér.

| primaryName | number of directed movies |
|---|---|
| Sam Newfield | 203 |
| Jesús Franco | 185 |
| Michael Curtiz | 155 |
| William Beaudine | 148 |
| Gilberto Martínez Solares | 142 |
| Lesley Selander | 133 |
| Lew Landers | 132 |
| Osman F. Seden | 130 |

V ďalšej tabuľke sa pozrieme na prvých 8 scenáristov„ zoradení podľa počtu napísaných scenárov na filmy, ktoré boli natočené.

| primaryName | number of written movies |
|---|---|
| William Shakespeare | 394 |
| Safa Önal | 288 |
| Erdogan Tünas | 261 |
| Bülent Oran | 251 |
| Kuang Ni | 212 |
| Jing Wong | 200 |
| Fernando Galiana | 171 |
| Jesús Franco | 167 |

Ďalej sme zanalyzovali, koľko existuje ľudí, ktorí film režírovali a zároveň napísali. Dokopy ich bolo  86 613. Prvých 5 aj s počtom tejto kombinácie vyzerá takto:

| primaryName | count |
|---|---|
| Jesús Franco | 151 |
| Jing Wong | 107 |
| Gilberto Martínez Solares | 88 |
| Mariano Ozores | 84 |
| Yôji Yamada | 84 |

## 3.2. Analýza populárnych a nepopulárnych filmov po doplnení nových dát

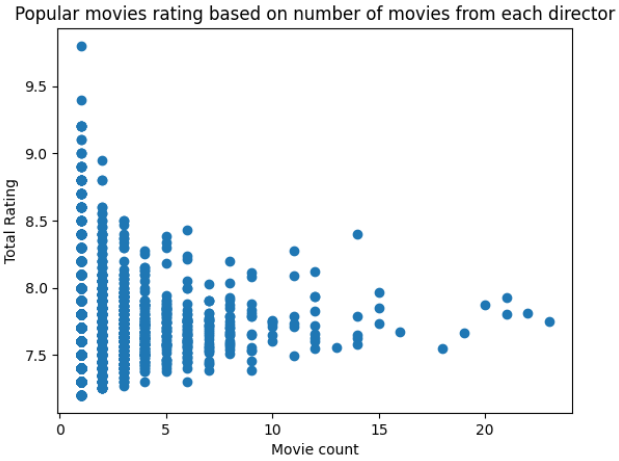
V nasledujúcej časti sme sa chceli zamerať na hlbšiu analýzu toho, čo môže robiť film populárnym a nepopulárnym.

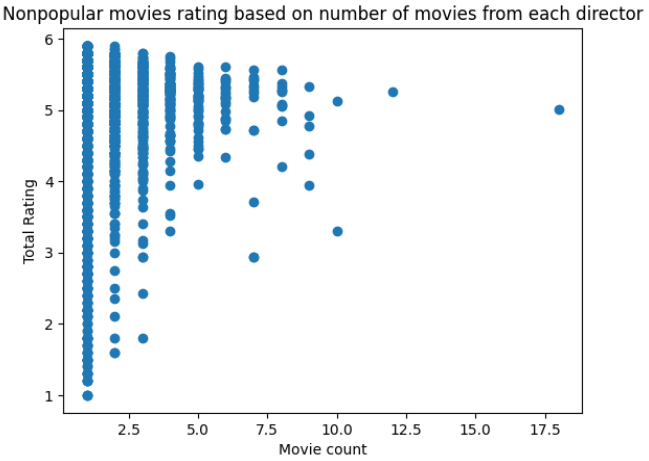
### Selekcia dát

Ako prvé sme si vyčlenili len tie dáta, ktoré majú počet hodnotení väčší ako priemerný počet hodnotení u všetkých filmov. Potom ich zoradili podľa ratingu a do populárnych sme zahrnuli prvú štvrtinu týchto filmov a do nepopulárnych štvrtú štvrtinu filmov. Tieto dve databázy sme následne pomocou scrapovania stránky imdb.com rozšírili o nové stĺpce:
'nudity',  'violence', 'profanity', 'alcohol', 'frightening', 'actors',
 'countriesOfOrigin', 'productionCompanies', 'budget', 'grossWorldwide'

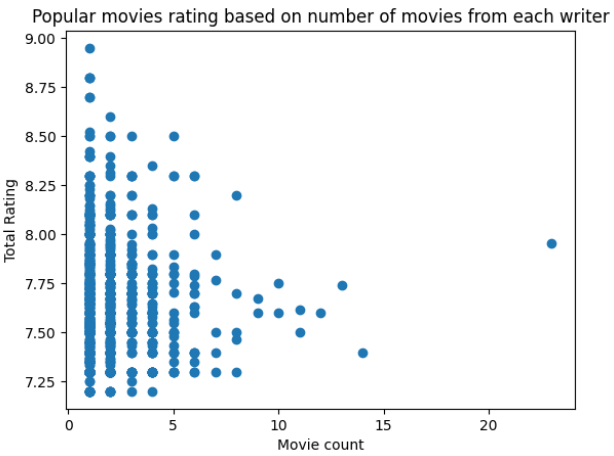
## Režiséri a scenáristi

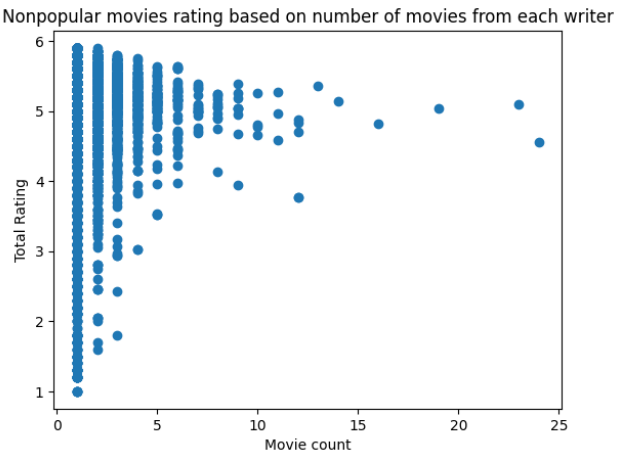
Zvlášť sme sa pozreli v tejto časti, na nami vyčlenené dáta pre obľúbené a pre neobľúbené filmy.




V grafe naľavo pozorujeme, že nie nutne s narastajúcim počtom filmov, ktoré jeden režisér vytvoril, rastie priemerné hodnotenie všetkých jeho filmov. Po vypočítaní korelácie dostávame silnú koreláciu a to 0,999. Taktiež v grafe na pravo nám vyšla korelácia 0,978, čo je opäť veľmi silná korelácia. Takže predpokladáme, že nejaký vzťah medzi týmito dvomi premennými môže existovať. Môžeme si všimnúť, že pri nepopulárnych filmoch režiséri nepokračovali v tvorení filmov, aj keď, tí, ktorí pokračovali dostali v priemere lepšie Hodnotenie, no stále zapadajú do nízkeho hodnotenia. V populárnych vidíme, že medzi úspešnými režisérmi sa našli častejšie takí, ktorí režírovali viac ako 5 filmov v porovnaní s Druhou tabuľkou.

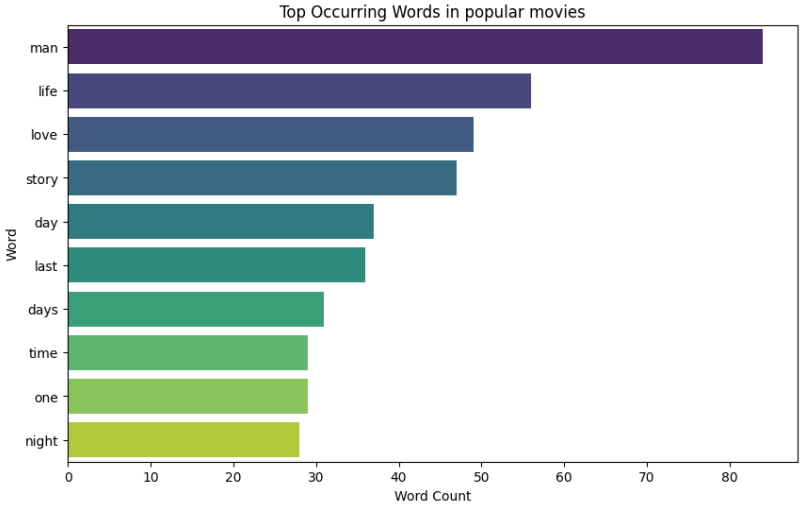


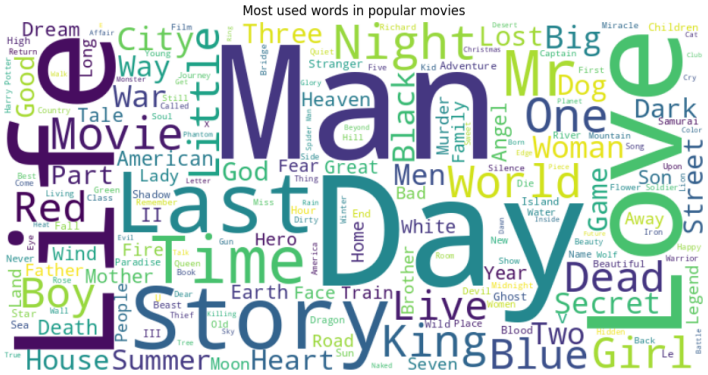




V tabuľkách s rovnakými dátami len tentoraz pre scenáristov opäť získavame podobnú distribúciu bodov. Pre nepopulárne filmy máme koreláciu 0,982, no pre populárne filmy Dostávame koreláciu iba 0,091.

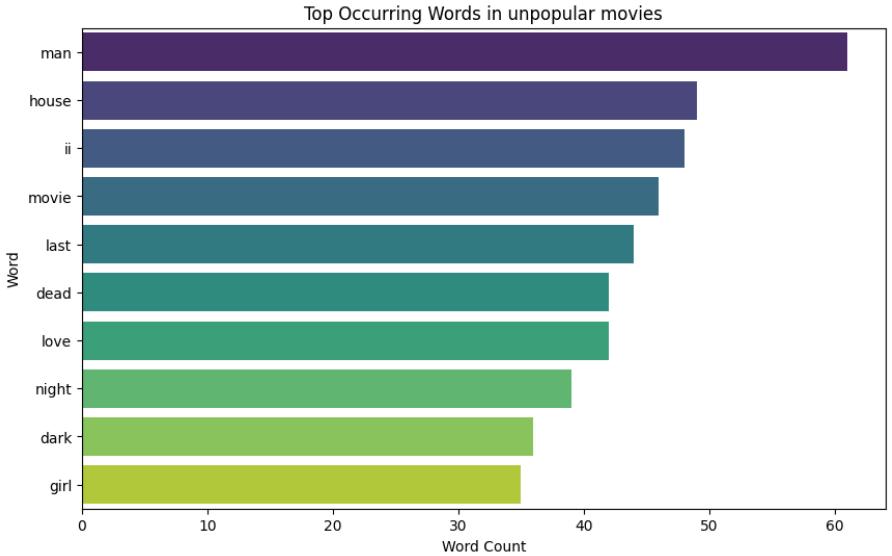
## Analýza názvov filmov

Frekvenciu slov sme získali pomocou CountVectorizera, vyfiltrovali sme ešte tzv. stop words - zámená, predložky, členy, aby nám vyšli len zmysluplné slová. Medzi najčastejšie vyskytujúce sa slová medzi populárnymi filmami patrili nasledovné. (vid. grafy nižšie)
Prvý graf ukazuje 10 najfrekventovanejších, ďalší mnohé s najväčším počtom výskytov.




Medzi najčastejšie vyskytujúce sa slová medzi nepopulárnymi filmami patrili nasledovné.
Prvý graf ukazuje 10 najfrekventovanejších, ďalší mnohé s najväčším počtom výskytov.




Slová z oboch kategórií sú pomerne všeobecné. Ale napríklad house v nepopulárnych sa však napr. môže viazať k nepopulárnym hororom, ktoré sa odohrávajú v nejakom dome hrôzy. (Horrorov s house v názve je vyše 600). Taktiež II môže znamenať nepodarené pokračovanie nejakého filmu.

## Logistická regresia

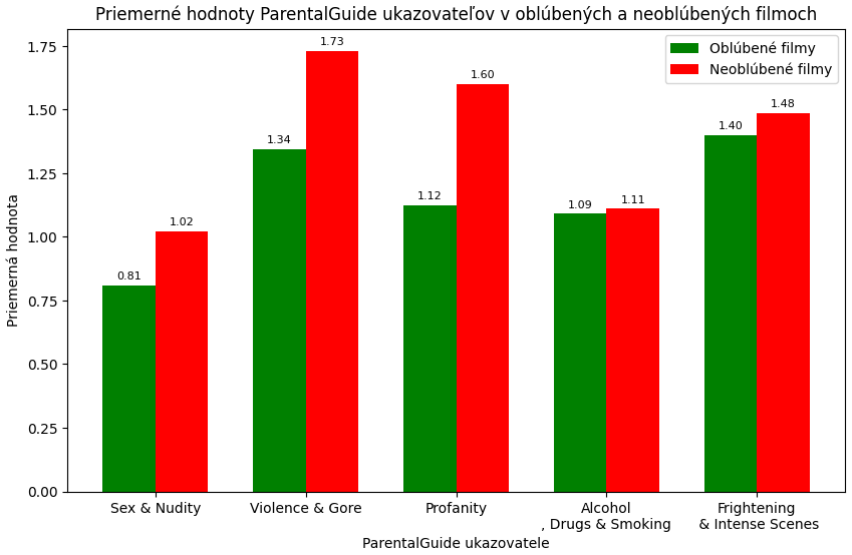
Vykonali sme aj jednoduchú logistickú regresiu, na oscrapovaných filmoch s doplňujúcimi informáciami, ktorá berie do úvahy všetky textové premenné, čo boli 'primaryTitle', 'genres', 'actors', 'productionCompanies','countriesOfOrigin', 'directors', 'writers'. Tieto všetky slová sme brali ako jeden celok viažuci sa k danému filmu. Ich konkatenáciou a následnou premenou na maticu pomocou CountVectorizera sme mohli spustiť logistickú regresiu, kde pozorovaná binárna premenná bola populárny. Kombináciou všetkých týchto faktorov sme na testovacej množine dosiahli F1 skóre 0.853, čo je pomerne vysoké číslo, takže model pomerne správne identifikuje pozitíva, s nízkym počtom falošných pozitív. Následným odobraním, vždy jedného textového stĺpca sme pozorovali ako sa F1 skóre mení. Najväčší rozdiel medzi F1 skórami vyšiel pri odobratí žánru, z toho možno usudzovať, že práve žáner má najväčší vplyv na obľúbenosť filmu. Pri ostatných stĺpcoch sme nebadali taký veľký rozdiel, z čoho možno usudzovať, že úspech má na svedomí viacero faktorov súčasne.

## Lineárna regresia

Vykonali sme aj lineárnu regresiu na normalizovaných číselných premenných - žánroch (ako dummy variables), roku vydania, čase v minútach, budgete, zárobkoch a iných. Napr. pri parental guides premenných sme zvolili ordinálnu premennú: 0 pre žiadne nevhodné scény daného typu (napr. alkohol , nahota), 1 pre jemné, 2 pre mierne, 3 pre drsné. R-squared vyšiel 0.538, takže náš model by sa dal ešte zlepšovať. Jednotlivé koeficienty vyšli najzaujímavejšie pri čase v minútach 0.34, počte hlasov 0.51, žánri Drama 0.26, budgete -0.16, roku vydania. -0.31. Je zaujímavé, že budget má záporný koeficient. Taktiež sme objavili kandidáta na najobľúbenejší žáner - dráma a na najhorší žáner horror, ktorý mal koeficient -0.16.

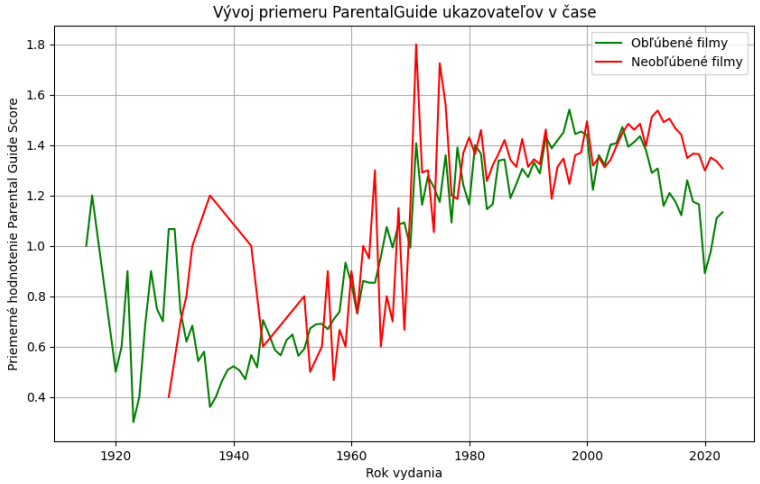
## Analýza hodnôt Sex & Nudity, Violence & Gore, Profanity, Alcohol, Drugs & Smoking, Frightening & Intense Scenes

Ako prvé sme si zistili priemery jednotlivých premapovaných hodnôt ParentGuide ukazovateľov pre obľúbené a neobľúbené filmy a odstránili riadky ktoré obsahovali None hodnotu.

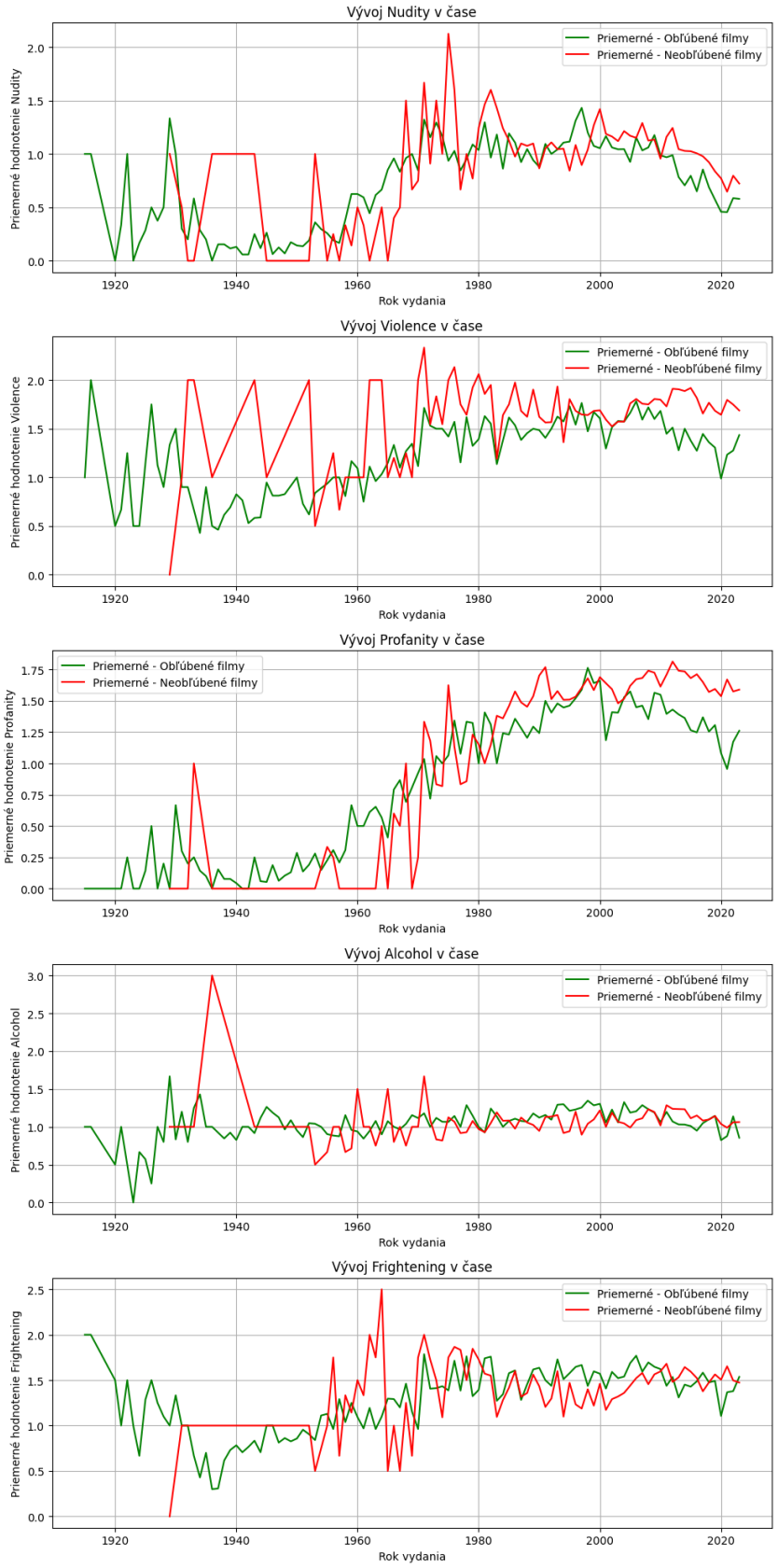



Z tohto grafu môžeme pozorovať, že najväčší rozdiel je v ukazovateľoch Violence & Gore a Profanity, z čoho môžeme dedukovať, že v nepopulárnych filmoch je o čosi viac násilia a nadávok než v populárnych filmoch.

Ďalej sme sa pozreli na priebeh priemeru všetkých ukazovateľov v populárnych a nepopulárnych filmoch v čase:

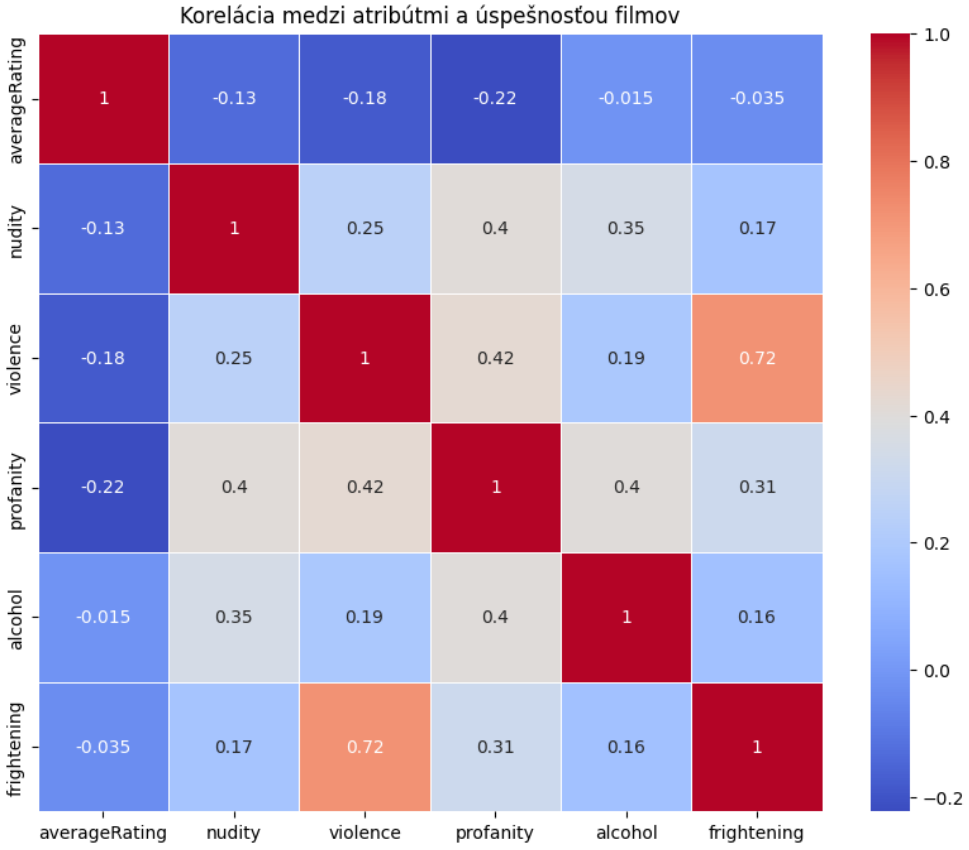



A následne sme sa pozreli aj na priebeh jednotlivých ukazovateľov v čase. Ako môžeme vidieť násilie v neobľúbených filmoch je vo väčšine času vyššie než v obľúbených no pri oboch prípadoch sa skoro po celé časové obdobie nachádza vo filmových scénach narozdiel od nadávok, ktoré sa postupom času viac začali používať vo filmoch.




Taktiež sme si vytvorili aj korelačnú maticu medzi priemerným hodnotením a parentalGuide ukazovateľmi obľúbených a neobľúbených filmov, z ktorej môžeme usúdiť, že niektoré parentalGuide ukazovatele so sebou navzájom súvisia ako napríklad násilne s desivými

scenármi. Taktiež skoro všetky majú záporný korelačný koeficient s priemerným hodnotením, čo môže naznačovať„ že obľúbenejšie filmy majú menšie parentalGuide skóre (je v nich menej nevhodných záberov).




To nám potvrdil aj štatistický test Mann-Whitney, s ktorým sme zistili, že filmy s nadpriemernou hodnotou jednotlivých ukazovateľov naznačujú, že ide o nepopulárny film keďže p-value pri tejto hypotéze vyšlo pri všetkých ukazovateľoch takmer 0.

## Analýza budgetu a zárobkov

Pomocou štatistického testu Mann-Whitney chceme odhadnúť, či filmy s nadpriemerným budgetom, resp. zárobkom majú vyššiu distribúciu ako podpriemerné.

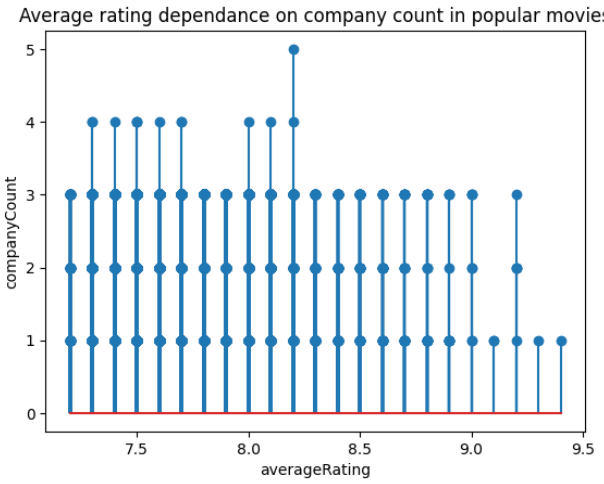
Pre budget sa nulová hypotéza nezamietla, práve naopak - p-hodnota vyšla 0.89 . Takže vyšší budget pravdepodobne vôbec nemusí zaručovať úspech.

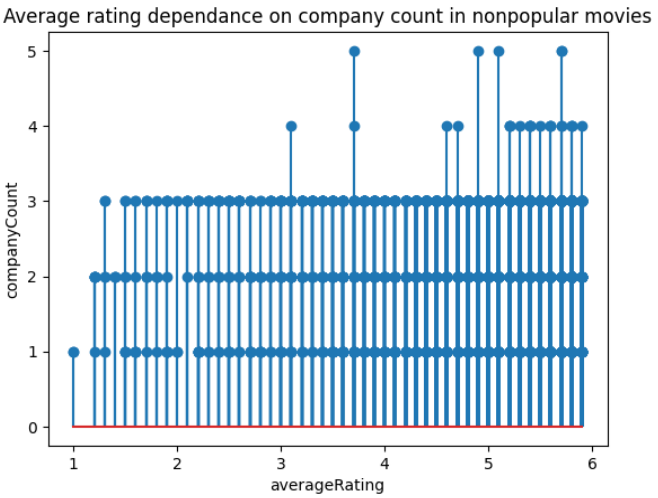
Pre zárobok vyšla p-hodnota takmer 0, takže dobrý zárobok pravdepodobne zaručuje aj kvalitu filmu.

Čistý zisk filmu sme odhadli ako zárobok-budget. Aj pre túto veličinu sme urobili štatistický test, pre rôznosť distribúcií. Pre čistý zisk vyšla ešte bližšie p-hodnota k 0 ako pri zárobku.

## Analýza produkčných spoločností

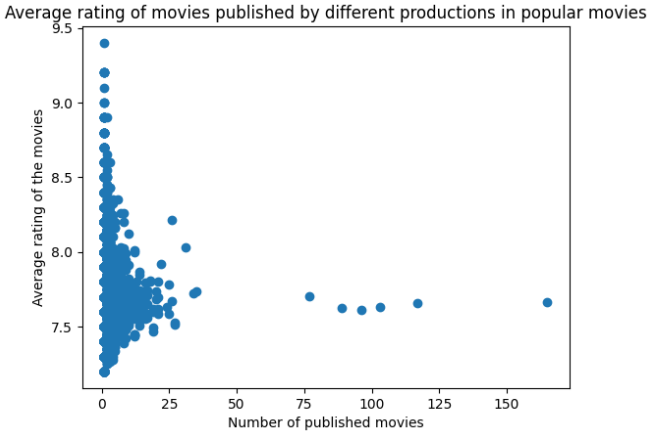
V nasledujúcich grafoch skúmame vplyv počtu rôznych produkčných spoločností na jednom filme na priemerné hodnotenie dané divákmi. V populárnych filmoch, aj v nepopulárnych filmoch sme vypočítali koreláciu medzi týmito dvomi premennými a v oboch prípadoch nám vyšla slabá korelácia. Pre populárne to je -0,114 a pre nepopulárne 0,104.

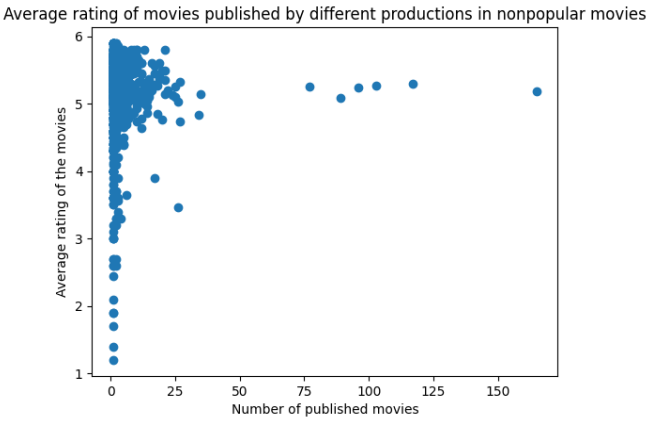



Z tohto a zároveň aj z distribúcie bodov (ktorá pôsobí ako nejaký zhluk bodov) na oboch grafoch nám vyplýva, že počet produkčných spoločností na jednom filme nemá vplyv na to, ako film bude obľúbený pre divákov.

Nasledujúce dva grafy ukazujú pre populárne a nepopulárne filmy distribúciu dvoch atribútov: priemerné hodnotenie priemerných hodnotení filmov pre každú produkciu (produkčnú spoločnosť) a počet vydaných filmov každej produkcie. V oboch grafoch ide vidieť, že väčšina produkcií vyprodukovala do 25 filmov.




Taktiež, distribúcia je veľmi podobná v obidvoch datasetoch, s rozdielom, že hlavný klaster bodov pre populárne filmy je v ľavom dolnom rohu a hlavný klaster pre nepopulárne filmy

je v ľavom hornom rohu. To čo si môžme všimnúť je, že produkcie, ktoré vyprodukovali do 10 filmov majú lepšie hodnotenia. Tento jav môže byť ovplyvnený tým, že bol použitý priemer pre priemerné hodnotenia, čo môže, pri viacerých dátach, spraviť väčšie rozdiely, keďže priemer je citlivý na outlierov.

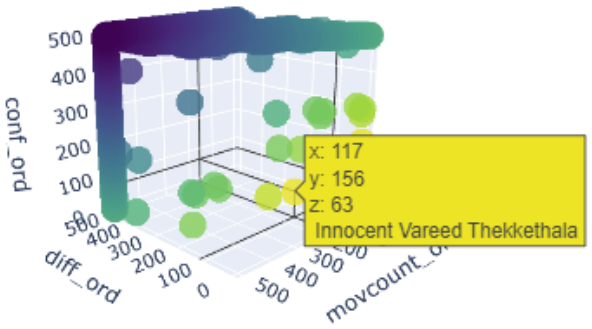
Teda nevieme tvrdiť, že pri vyšších počtoch vyprodukovaných filmov produkčnými spoločnosťami, je očakávané vyššie priemerné hodnotenie.

## Analýza obľúbených hercov

Rozhodli sme sa zanalyzovať aj najlepších a najobľúbenejších hercov. Keďže tieto údaje máme len v nových tabuľkách v konečnom výsledku nevieme povedať či má prítomnosť konkrétnych hercov vplyv na obľúbenosť filmov, našou analýzou však prichádzame na iné zaujímavé podnety.
Urobili sme si 3 oddelené tabuľky:

- top500actors_diff: táto tabuľka zobrazuje 500 takých hercov, ktorých rozdiel počtu populárnych filmov v ktorých hrali a počtu nepopulárnych filmov, v ktorých hrali je čo najväčší
- top500actors_conf: v tejto tabuľke sú herci iba tí ktorí sú aj v tabuľke populárnych filmov a zoradený podľa dolnej hranice jednostranného 90%-ého intervalu spoľahlivosti, ktorý počíta vhodnú závislosť medzi počtom populárnych filmov daného herca, priemerom hodnotení týchto filmov a smerodajnou odchýlkou tohoto hodnotenia
- top500actors_movcount: táto tabuľka ukazuje jednoducho 500 hercov s najväčším počtom populárnych filmov v ktorých hrali

Poďme si vysvetliť nasledujúci graf. Jeden bod v grafe zodpovedá jednému hercovi. Jednotlivé osi ukazujú poradie herca v tabuľkách vyššie. Ak sa herec aspoň v jednej z tabuliek nenachádzal, priradili sme mu hodnotu 501. Farba ukazuje pomer obľúbenosti, ktorý je vypočítaný závislosťou poradia v tabuľkách vyššie. Čím viac sa farba približuje k žltej, tým herca považujeme za obľúbeného. Tabuľke s intervalmi spoľahlivosti sme dali najväčšiu váhu, lebo ju považujeme za dôveryhodnú.

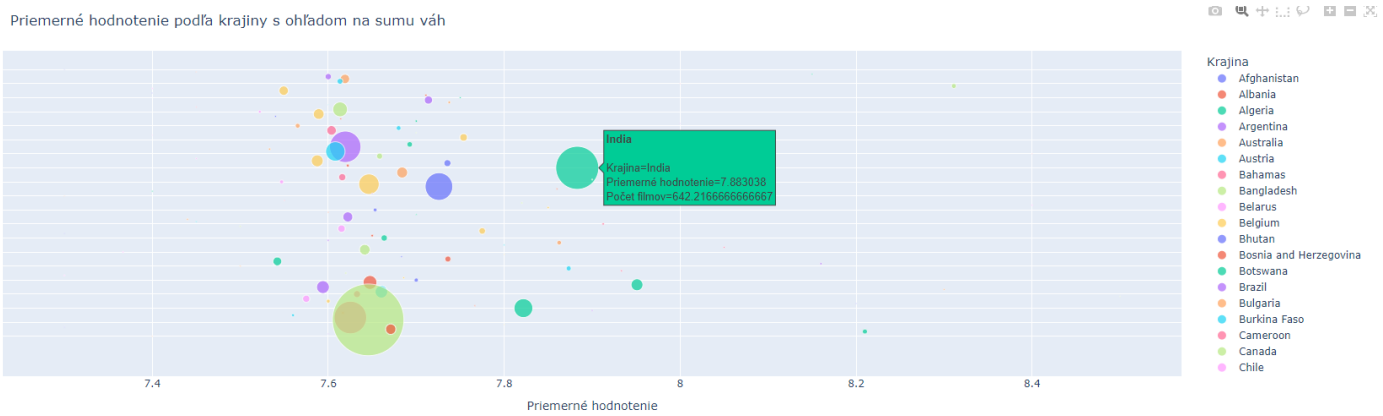


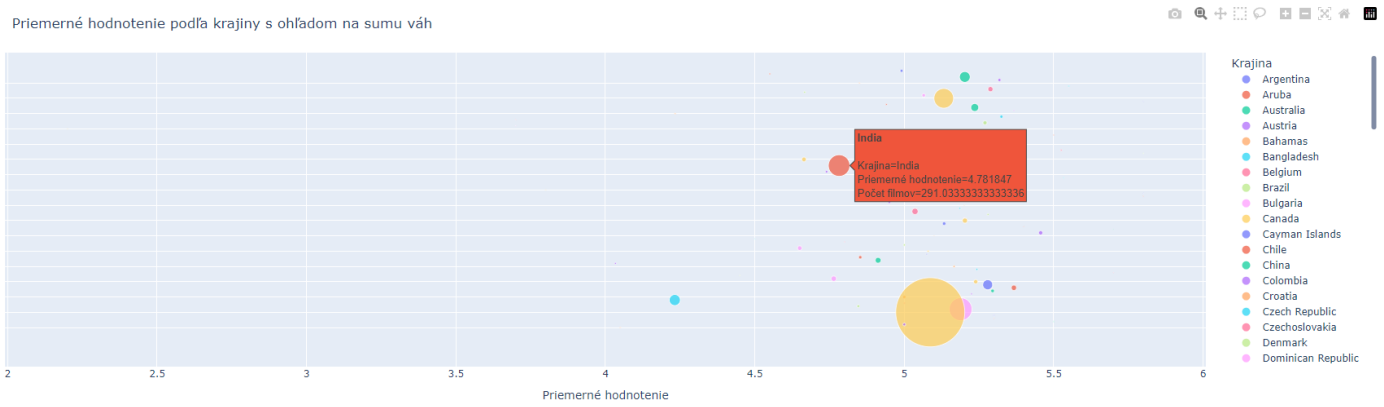

Touto analýzou dostávame v tabuľkách veľa hercov, ktorých veľmi nepoznáme. Takáto analýza nemá priamy vplyv na obľúbenosť filmov, ale je zaujímavé vidieť akí rôzni herci sú

na vysokých priečkach našej analýzy. K lepšej analýze by sa nám oplatilo mať hercov v kompletnej tabuľke movies a nie len pri populárnych alebo nepopulárnych. Tiež by bolo zaujímavé preskúmať súvis medzi ratingom, hercom a počtom minút ktoré herec vo filme účinkoval. Predpokladáme, že tu by sme dostali smerodajnejšie výsledky.

## Vplyv krajiny na hodnotenie filmu

V našej databáze sa nachádza krajina alebo krajiny pôvodu, z ktorých film pochádza. Do pomeru sme dali hodnotenie a krajnu pričom sme vzali do úvahy fakt, že počet filmov môže mať tiež vplyv (ak mal film viacero krajín pôvodu, dali sme vážený priemer). Ak sa pozrieme na najviac neobľúbené krajiny s väčším počtom filmov, aby to nebolo len o náhode (najviac naľavo a čo najväčší bod) a porovnáme ich s populárnymi (v grafe najviac napravo, tiež čo najväčšie body)  tak nájdeme veľa rovnakých krajín a teda krajina by nemala mať veľký vplyv na hodnotenie, keďže sa nachádza v oboch grafoch s podobným počtom filmov, ale v oboch extrémoch s vysokým aj nízkym hodnotením.



# 4. Scrapovanie stránky imdb.com

Na parsovanie html kódu imdb.com sme použili nasledujúce knižnice: requests - (vykonanie HTTP požiadaviek na získanie obsahu webových stránok), BeautifulSoup z bs4 (extrahovanie dát zo stránok), selenium (ovládanie webového prehliadača - v tomto prípade ChromeDriver na dynamické načítanie stránok a získavanie dát) a webdriver_manager.chrome (automatická inštalácia a riadenie ChromeDrivera).
Vytvorili sme si dve funkcie:

1. 'parentalguide_data' - v ktorej sme ziskavali dáta z adresy f'https://www.imdb.com/title/{id}/parentalguide' pričom hodnota id bola hodnota 'tconst' už zo získanej databázy. Následne sme uložili cez jedinečné triedy hodnoty Parents Guide - Sex & Nudity, Violence & Gore, Profanity, Alcohol, Drugs & Smoking, Frightening & Intense Scenes
2. 'other_data' - taktiež sme si v tejto funkcii získali cez jedinečné triedy dáta - Actors, Country of origin, Production companies, Budget a Gross worldwide, tentokrát z adresy f'https://www.imdb.com/title/{id}/', ktorú sme si museli vždy dynamicky načítať aby sa nám uložil html kód. To zabralo aspoň 11 sekúnd pre každý film, preto sme si museli spuštenia tohto scriptu rozdeliť. Taktiež sme si museli dávať pozor aby sa nám ukladali hodnoty v rovnakej mene - čo sme ohraničili tým, že sme doláre a libry prepočítali na eurá a ostatné mená nezahrnuli do výstupných dát.

Celý kód sa nachádza v súbore script.py

# Záver

Napriek tomu, že výber obľúbeného filmu je subjektívna vec každého z nás, v tejto práci sme prišli na viaceré faktory, ktoré vedia naznačiť či je film objektívne obľúbený alebo nie. Poďme zodpovedať otázky, ktoré sme si položili na začiatku.

Výber vhodného  žánru filmu ovplyvňuje jeho obľúbenosť. Medzi obľúbené voľby patrí Dráma, Dokument alebo Komédia. Neobľúbenými sú filmy kde sú skombinované rôzne žánre, ktoré majú málo spoločného. Čo nás zaujalo je, že žáner Sci-fi patrí k tým menej obľúbeným. V logistickej regresií nám vyšlo, že práve žáner je najdôležitejšou nečíselnou premennou.

Čo sa týka dĺžky filmu, najkvalitnejšie boli filmy s dĺžkou 20-40 minút. Celovečerné filmy s klasickou dĺžkou jeden a pol hodiny mali najnižší podiel kvalitných filmov. Tiež naša analýza ukázala, že filmy s dĺžkou filmu nad 120 minút často patria medzi obľúbenejšie.

Berúc do úvahy rok vzniku,  filmy medzi vojnami nepatrili medzi kvalitné, no môžeme povedať, že od konca druhej svetovej vojny, filmy naberajú na kvalite až do súčasnosti.

Zaujímavé zistenie pre nás bolo, že vyšší rozpočet neznamená, že je film kvalitnejši. Táto hypotéza, že vyšší rozpočet znamená kvalitnejší film sa nám nepotvrdila. Mohli sme to pozorovať aj v grafe o vývoji rozpočtu v čase, ktorý zobrazoval že priemerný rozpočet narastal v dobe aj pre populárne aj pre nepopulárne filmy.

Množstvo násilia, zastrašovania, sexuality, nadávok alebo alkoholu ovplyvňuje obľúbenosť filmu, čím je ich tam viac má väčšiu pravdepodobnosť byť neúspešným.

Taktiež sa ukázalo, že počet rôznych produkcii na jedom filme nemá nejaký vplyv na hodnotenie filmov. Zároveň sme zistili, že ani kvantita filmov pre rôzne produkcie nemusí nutne prinášať lepšie hodnotenia.

Za relevantné faktory určujúce obľúbený alebo kvalitný film považujeme žáner, výnos, počet hodnotení, dĺžku filmu, rodičovské pomôcky a tieto hodnoty vo vyššie opísaných súvislostiach.

V budúcnosti by sa dalo na nadstavbe popracovať hlavne na zoscrapovaní a uložení hodnôt pre všetky filmy, nie len pre nami vybrané obľúbené a neobľúbené, čo by nám pomohlo aplikovať hlbšiu a presnejšiu analýzu. Avšak spustenie takého kódu na 50000 filmoch by mohlo zabrať aj celý týždeň. :)