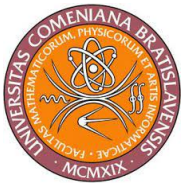
Čo robí filmy obľúbenými a neobľúbenými?

Dátová veda: tím: GOATs
P. Franček, L. Jankola, J. Kamas, A. Pauer, D. Števaňák



Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
December 2023

Obsah

0.1 Úvod do témy

Rozhodli sme sa pre tému obľúbenosti filmov, pretože sú u väčšiny ľudí neodmysliteľnou súčasťou života. Ľudia vnímajú rôzne filmy inak. Niektoré pozerajú radi a niektoré nie. Cieľom projektu je preskúmať základné trendy v dátach a pozorovať, ktoré faktory majú vplyv na obľúbenosť filmov. Význam projektu spočíva v spracovaní dát a tvorbe grafov, ktoré umožňujú všímať si rôzne súvislosti. Práca sa zameriava na faktory ako žánre, typ filmu, dĺžka, režiséri,no skúma aj koreláciu vybraných dát. V projekte sme použili metódy ako analýza hlavných komponentov, zhlukovanie, regresná analýza a taktiež sme testovali štatistické hypotézy.

0.2 Popis dát

V projekte sú použité dáta od spoločnosti IMDB, ktoré boli stiahnuté z oficiálnej stránky https://datasets.imdbws.com/. Dáta sú rozdelené do viacerých súborov, každý zodpovedá inej charakteristike filmov (alebo personálu). Dáta popisujúce filmy sa nachádzajú v súboroch `title_...tsv` a majú spoločný atribút `tconst` (id filmu). Ostatné súbory `name_...tsv` obsahujú dáta o ľudoch, ktorí sa podieľali na tvorbe filmu. V projekte pracujeme najmä s atribútmi ako

- `titleType` – typ filmu (epizóda, seriál, šou, ...)
- `starYear` – rok vydania filmu
- `runtimeMinutes` – čas trvania filmu
- `averageRating` – priemerné vážené hodnotenie filmu
- `genre` – žáner filmu
- `numVotes` – počet hodnotení

Dáta boli príliš veľké na to, aby sme ich mohli uložiť na platformu GITHUB. Z dát sme teda vybrali podmnožinu, na ktorej sme následne vykonali analýzu. Z datasetu `title_ratings.tsv` sme zobrali každý desiaty film. Na základe toho sme upravili aj ostatné súbory. Tvorba nových dátových súborov sa nachádza v súbore `scrape.py`. Takto vytvorené dátové súbory sme nahrali na GITHUB, odkiaľ boli načítavané. Upravené dátové súbory sú

- `title_ratings_scraped.tsv` – hodnotenie filmov
- `title_basics_scraped.tsv` – základné charakteristiky ako rok vydania alebo trvanie filmu)
- `title_crew_scraped` – režiséri a spisovatelia pre daný film
- `title_principals_scraped.tsv` – ľudia, ktorí pracovali na filme a ich úlohy
- `title_akas_scraped.tsv` – jazyk v ktorom bol film natočený

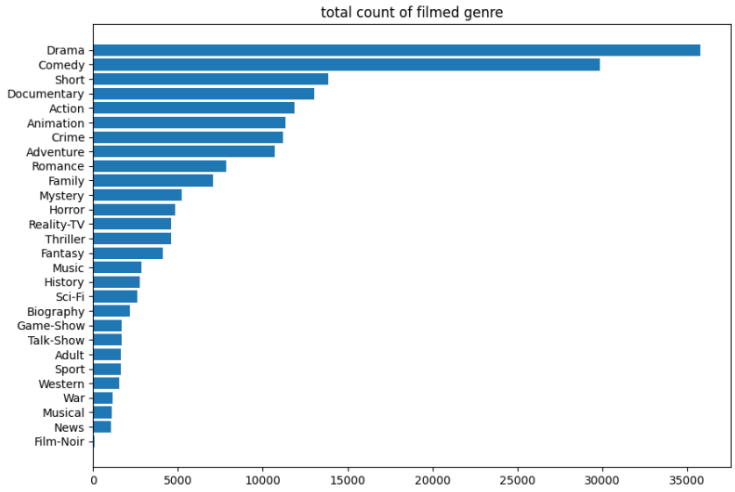
Kapitola 1

Vplyv žánrov na obľúbenosť

Skúmali sme, ktoré žánre majú najvyššie alebo najnižšie hodnotenie. Keďže film môže mať viacero žánrov, žánre budeme v datasete reprezentovať pomocou binárnych vektorov. Použili sme tzv. `multi - label encoding`.

1.0.1 Počet filmov a hodnotenie

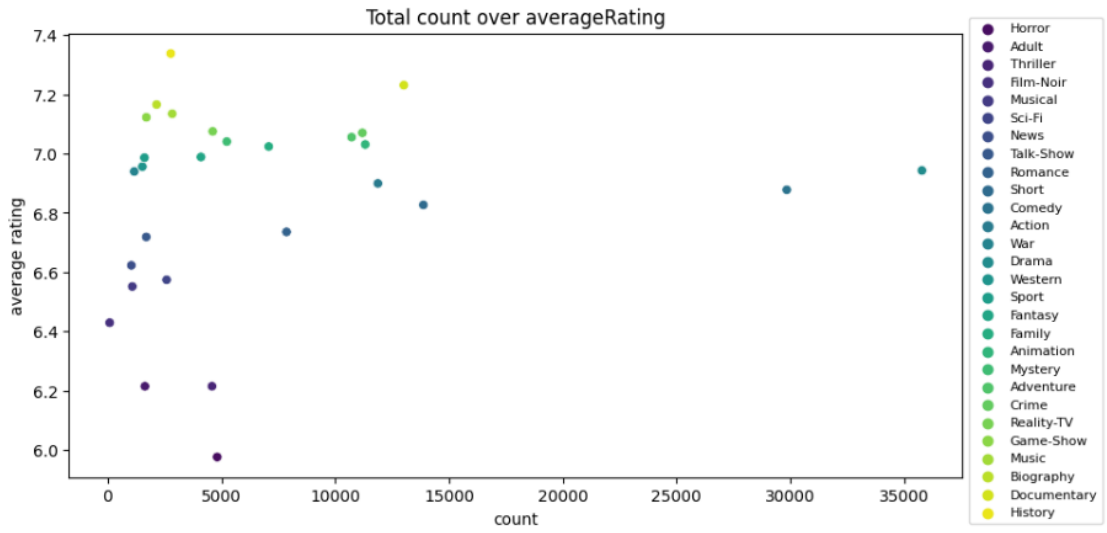
Intuícia nám hovorí, že populárne žánre sa natáčajú viac ako žánre, ktoré sú menej obľúbené. Spočítame teda celkový počet filmov prislúchajúcich ku každému žánru.



Obr. 1.1: Počet filmov pre žáner

Z obrázku 1.1 dostávame prvotný odhad, ktoré filmy by mohli byť obľúbené a ktoré nie. Najviac filmov prislúcha žánrom `Drama` a `Comedy`. Počet filmov týchto žánrov je oveľa väčší ako u ostatných žánrov. Domnievame sa teda, že tieto žánre budú dominovať medzi obľúbenými. Najmenej filmov prislúcha žánrom `Musical`, `News` a `Film-Noir`(čierno biely film).

Pozrime sa taktiež na priemerné hodnotenie filmov pre každý žáner a porovnajme s predošlými výsledkami.



Obr. 1.2: Závislosť hodnotenia od počtu filmov pre žánre

Obrázok 1.2 ukazuje, že aj žánre s malým počtom filmov v datasete majú vysoké priemerné hodnotenie hodnotenie. Žánre `Drama` a `Comedy`, ktoré sme odhadovali ako obľúbené majú stredné priemerné hodnotenie. To značí, že žánre s vysokým počtom filmov nemusia obľúbené. Taktiež žánre s najmenším počtom filmom nemajú najmenšie hodnotenie. Na obrázku môžeme vidieť aj porovnanie priemerného hodnotenia žánrov. Najvyššie hodnotenie majú žánre `Documentary`, `History`, `Biography` a najnižšie pre `Horror`, `Adult`, `Thriller`.

Môžeme teda povedať, že hodnotenie žánrov (teda aj obľúbenosť) nezávisia od počtu filmov prislúchajúcim ku jednotlivým žánrom.

1.1 Regresný Model

Dôležitosť jednotlivých žánrov teraz porovnáme pomocou lineárnej regresie.
Model : $\texttt{averageRating} = \beta_0 + \beta_i\ \texttt{žáner}_i + \epsilon, \epsilon \sim N(0, \sigma^2), \quad i = 1, ..., 29$
Keďže pozorujeme dôležitosť atribútov v modeli normalizujeme dáta na `z-score`. Koeficienty v modeli vyšli nasledovne:

Tabuľka 1.1: Koeficiety v regresnom modeli

genre	coef	pvalue
Documentary	0.154962	0.000000e+00
Drama	0.112811	1.437973e-183
Comedy	0.071277	9.067108e-79
Crime	0.064295	7.247770e-77
Reality-TV	0.064093	2.228630e-73
Animation	0.059778	4.534965e-61
Adventure	0.055612	6.880141e-51
Fantasy	0.048036	4.307523e-52
History	0.047268	3.221923e-48
Mystery	0.046537	1.256989e-44
Music	0.039468	1.027506e-35
Short	0.037932	1.069155e-28
Family	0.034661	5.366196e-27
Western	0.032337	5.137298e-24
Game-Show	0.028572	6.368496e-18
Sport	0.021406	1.157950e-11
Action	0.010073	6.005625e-03
Talk-Show	0.009011	1.015444e-02
Biography	0.007645	1.742116e-02
War	0.004151	1.860162e-01
Sci-Fi	-0.008714	5.568331e-03
Film-Noir	-0.012034	1.049308e-04
Romance	-0.014514	7.458417e-06
Musical	-0.016101	2.391968e-07
News	-0.016666	1.145900e-06
Adult	-0.032924	1.454174e-24
Thriller	-0.065821	2.954932e-91
Horror	-0.115435	4.926312e-273

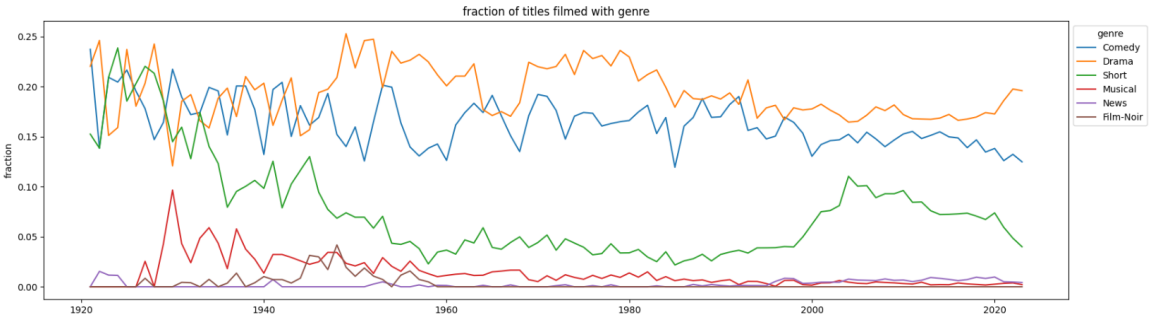
Keďže väčšina koeficientov v modeli je blízka k nule, nevieme povedať, či sú alebo nie sú v modeli významné, a preto sme urobili aj štatistický test. Týmto testom testujeme, či sa koeficienty v skutočnosti rovnajú nule. Hypotézy sú konštruované nasledovne : $H_0 : \beta_i = 0$ a $H_1 : \beta_i \neq 0$. Keďže p hodnota je menšia ako 0.05 pre všetky koeficienty okrem `War`, H_0 zamietame. To značí, že takmer všetky koeficienty sú v modeli významné a ich analýza je zmysluplná.

Najväčšie koeficienty vyšli pre žánre `Documentary`, `Drama`, `Comedy`. To znamená, že ak film má práve takýto žáner, odhad pre jeho hodnotenie sa najviac zvýši. Môžeme povedať, že tieto žánre majú vplyv na obľúbenosť a patria medzi obľúbené. Všimnime si, že najvplyvnejšie žánre sú podobné ako v predošlých analýzach.V modeli taktiež vystupujú záporné koeficienty. Tie odhad pre hodnotenie filmu znižujú. Na základe toho môžeme povedať, že tieto žánre robia filmy skôr neobľúbenými. Sú to žánre `Adult`,`Thriller` a `Horror`.

1.2 Rok vydania filmu a hodnotenie žánrov

Výsledky analýzy v časti 1 nepotvrdili vplyv počtu filmov na hodnotenie. Príčinou môže byť aspekt, že v minulosti boli populárne iné žánre ako teraz. Rozhodli sme sa teda preskúmať ako sa mení počet

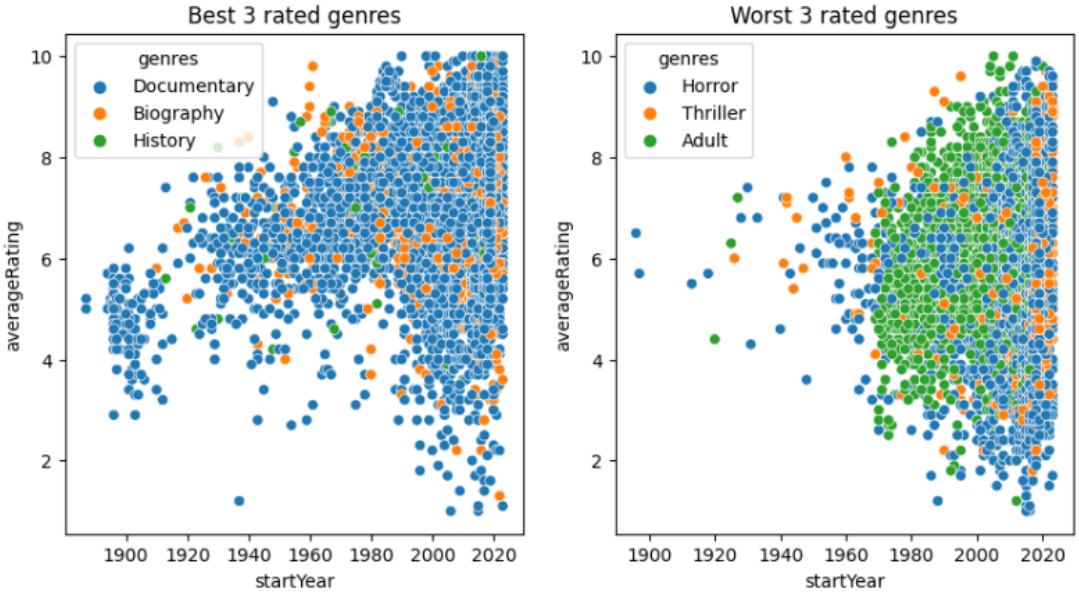
filmov pre žánre s najväčším(najmenším) počtom filmov vzhľadom na rok.



Obr. 1.3: Podiel počtu filmovaných žánrov

Vidíme, že vývoj počtu filmov pre žánre na obrázku má stabilný priebeh. Žánre `Comedy` a aj `Drama` si udržiavajú filmovanosť okolo hodnoty 20 %. Naopak `News` a `Film-Noir` sú systematicky menej filmované. Jedinú väčšiu zmenu vidíme pre žáner `Short`, ktorého filmovanosť postupne klesla. Domnievame sa teda, že rok nemá významný vplyv na filmovanosť jednotlivých žánrov.

Preskúmali sme aj vplyv roku vydania filmu na hodnotenie pre žánre, ktoré boli najlepšie hodnotené.



Obr. 1.4: Hodnotenie vybraných žánrov vzhľadom na rok

U prvej skupiny žánrov sa zdá, že rok vydania koreluje s priemerným hodnotením. Naopak, pri druhej skupine žánrov nepozorujeme veľkú závislosť. Korelačné koeficienty pre jednotlivé skupiny sú $\rho_1 = 0.23$ a $\rho_2 = 0.01$. Teda hodnotenie pre najlepšie žánre koreluje s rokom vydania filmu. Možno to vysvetliť tak, že najlepšie hodnotené žánre sa postupne časom zlepšovali oproti najhoršie hodnoteným, ktoré sa udržiavali na približne rovnakom priemernom hodnotení.
Korelácia hodnotenia a roku vydania filmov v celom datasete je $\rho = 0.14$. Vzhľadom na počet dát

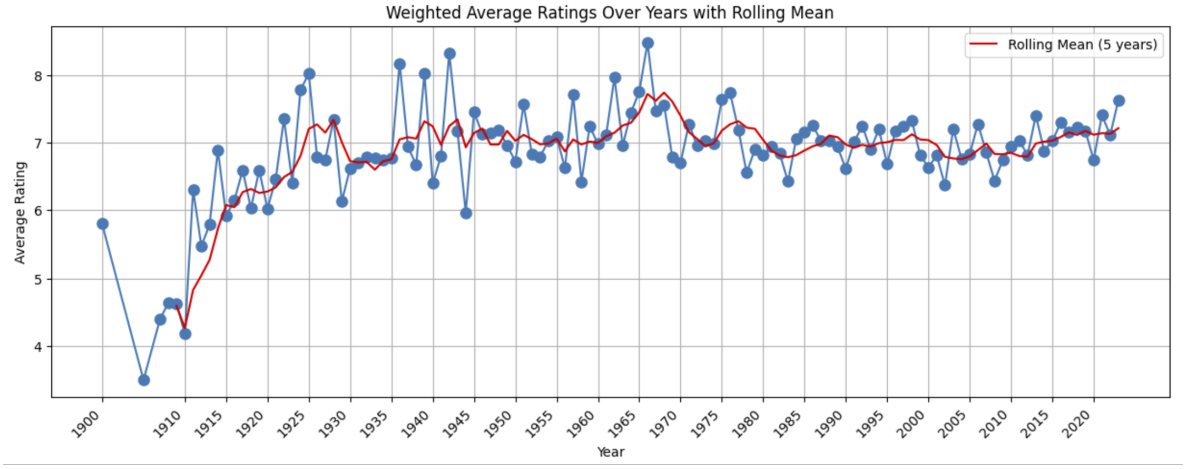
si myslíme, že závislosť medzi týmito premennými existuje a obrázok 1.4 potvrdzuje, že rok vydania filmu má vplyv na hodnotenie žánrov a teda aj na obľúbenosť.

1.3 Rok vydania a hodnotenie filmov

Skúmali sme taktiež ako sa počas rokov pohybovalo priemerné hodnotenie. Keďže také typy ako krátke filmy, video hry, mini série, videá či televízne špeciály sú často málo hodnotené a zriedka ich ľudia vôbec pozerajú tak sme ich pre našu potrebu vyhodili. Odfiltrovali sme tie, ktoré sú ohodnotené a spravili vážený priemer za každý rok pomocou počtu hodnotení. Čiže tie filmy, ktoré mali viac hodnotení celkovo zaváži viac ako tie s menej hodnoteniami. Následne sme použili vážený priemer, aby sme lepšie videli trendy v dátach.

Na grafe 1.5 môžeme vidieť, že ten trend priemerného hodnotenia sa ustálil niekde okolo roku 1925 na hodnote 7.0. Avšak to že predtým bol menší je pravdepodobne kvôli tomu, že sa toľko nenatáčalo a nehodnotilo tým pádom bola vzorka veľmi malá a neinformatívna.

Najzaujímavejšie je teda pozerať sa na roky kedy nastal nejaký skok. To môžeme vidieť v rokoch 1965-1975. Je však ťažké polemizovať čím to mohlo byť spôsobené, no jednou z viacerých možností je, že počet televízií v domácnostiach rapídne narástol a teda aj miera pozerania. Taktiež sú tieto roky mnohými považované za éru kedy boli vydané najkultovejšie filmy ako Čeľuste, Krstný otec a iné.



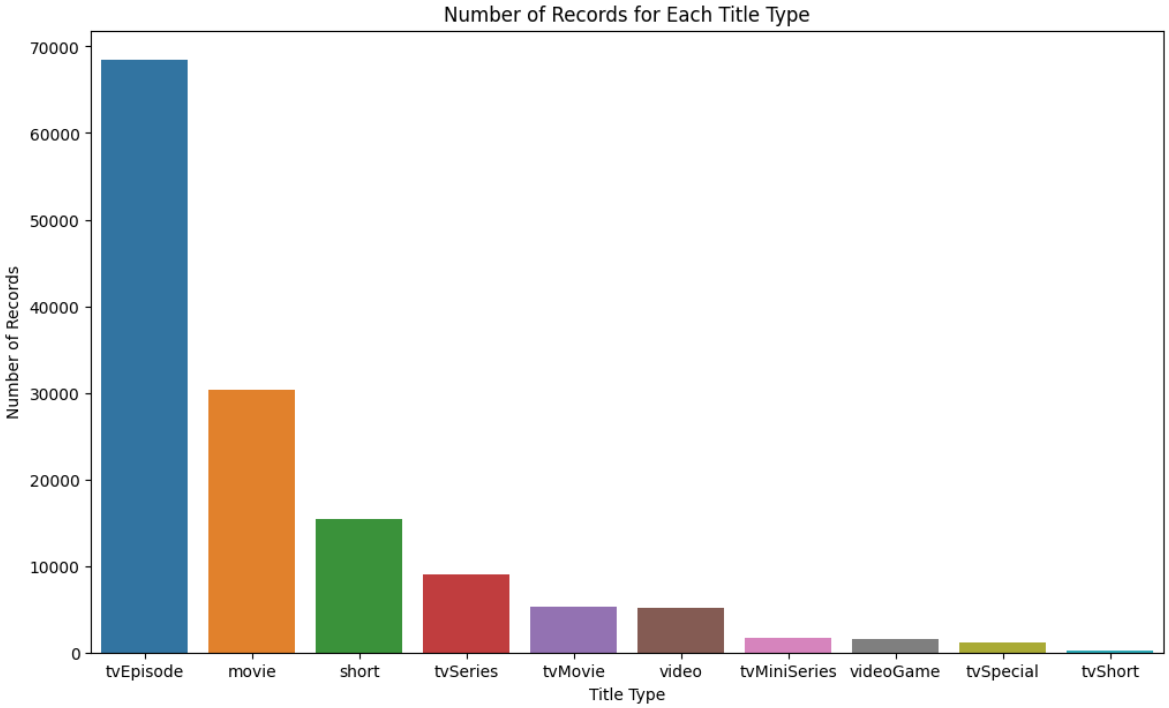
Obr. 1.5: Vážený priemer hodnotení počas rokov s kĺzavým priemerom

Kapitola 2

Vplyv typu filmu na obľúbenosť

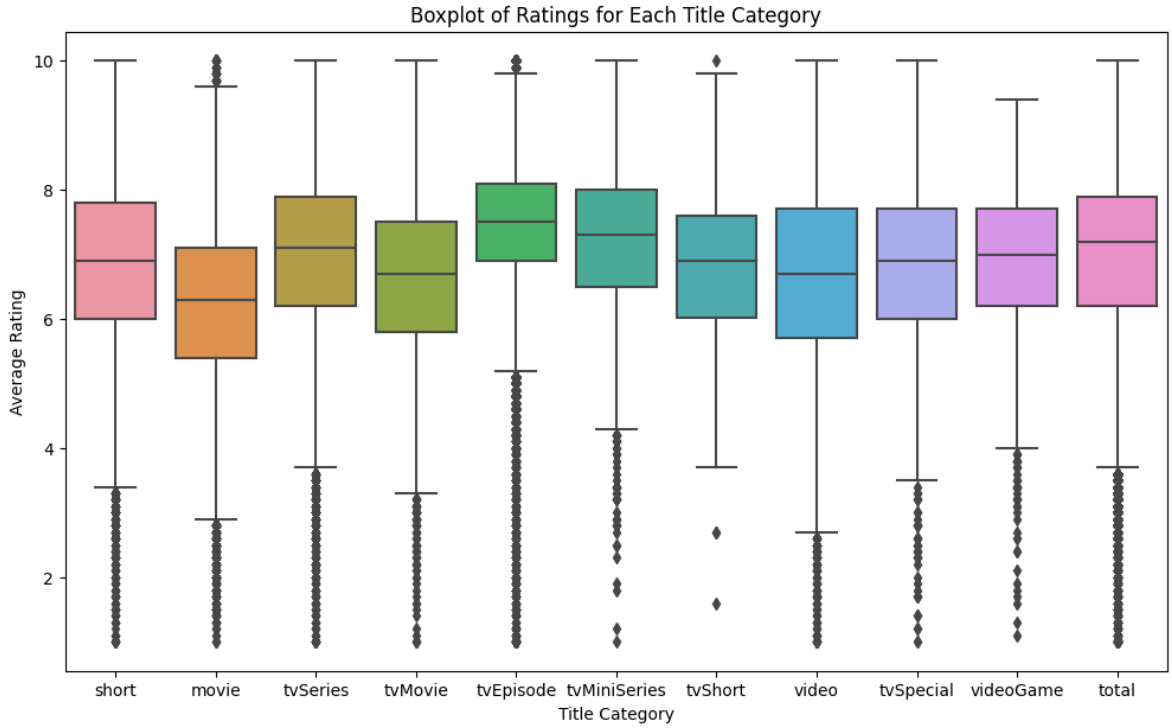
Ďalším zaujímavým parametrom je typ filmu. V našich dátach je označený ako `titleType`. Jednotlivé typy filmu sú nasledovné:

- `movie`
- `short`
- `tvEpisode`
- `tvMiniSeries`
- `tvMovie`
- `tvShort`
- `tvSpecial`
- `video`
- `videoGame`



Obr. 2.1: Počet záznamov pre každý typ filmu

Pri tomto kategorickom parametri sme chceli zistiť, či niektorý z typov má významne odlišné hodnotenie oproti spoločnému priemeru. Na vykreslenie sme použili boxploty, ktoré prehľadne ukazujú nie len priemer ale aj rozptyl a kvartily.



Obr. 2.2: Boxploty pre hodnotenie typov filmov

Už z Obr. 2.2 vidno, že hodnotenie pre `movie` je v priemere výrazne nižšie. Pre jednoznačnejšie určenie, či niektorý typ filmu dostáva odlišné hodnotenie sme sa rozhodli použiť korelačnú analýzu. Vytvorili sme si preto pomocnú tabuľku ktorá obsahovala nasledovné stĺpce: `averageRating`, a pre každý typ filmu príslušný binárny stĺpec obsahujúci 1 ak ten film je daného typu inak 0. Potom sme vyrátali Spearmanov korelačný koeficient a výsledky sú nasledovné:

type	**Correlation**
movie	-0.2959
short	-0.0295
tvEpisode	0.3154
tvMiniSeries	0.0163
tvMovie	-0.0496
tvSeries	-0.0173
tvShort	-0.0063
tvSpecial	-0.0152
video	-0.0591
videoGame	-0.0106

Tabuľka 2.1: Výsledky Spearmanovho korelačného testu

Významná negatívna korelácia sa ukázala pre `movie` a to -0.2959. To potvrdzuje odhady pozorované na boxplotoch. Zároveň vidno značnú pozitívnu koreláciu pri `tvEpisode` a to 0.3154.

Na záver je vhodné ešte otestovať koreláciu korelačným testom. Ten však zamietne nulovú hypotézu veľmi jednoducho, keďže máme pomerne veľké dáta. V nasledújúcich tabuľkách sú zobrazené typy, pri ktorých bola zamietnutá nulová hypotéza $H_0 = \rho < 0$ resp. $H_0 = \rho > 0$.

type	**statistic**	**pvalue**
tvEpisode	0.331513	0.000000e+00
tvMiniSeries	0.015841	1.937240e-09

type	**statistic**	**pvalue**
movie	-0.311289	0.000000e+00
short	-0.041122	4.169885e-53
tvMovie	-0.058351	7.972035e-105
tvSeries	-0.006783	5.838957e-03
tvShort	-0.007516	2.601841e-03
tvSpecial	-0.014187	6.653041e-08
video	-0.051743	7.168502e-83
videoGame	-0.013745	1.608362e-07

Tabuľka 2.2: Výsledky Spearmanovho korelačného testu

Kapitola 3

PCA a Zhlukovanie

Jednou z možností ako sa pozrieť na jednotlivé atribúty, ktoré vplývajú na lepšie hodnotenie filmov je snaha zoskupiť podobné filmy do zhlukov na základe podobnosti ich atribútov, pozrieť sa potom na zhluky s vysokým hodnotením a preskúmať ich alebo hľadať spoločné črty. Avšak pracovať so všetkými dátami bolo pre zhlukovanie pamäťovo náročné, a teda sme potrebovali vybrať menšiu vzorku (zároveň aj kvôli veľkému počtu chýbajúcich hodnôt v celom datasete). Rozhodli sme sa zvoliť filmy ktoré mali najvyšší počet hodnotení, keďže tieto vyvolali najväčší ohlas a analyzovať práve ich vplyv - taktiež ich počet chýbajúcich hodnôt sa pohybuje pod 1%. Počet filmov, ktoré sme zvolili bol 1000 - nasledovnú analýzu sme skúšali spúšťať aj pre výber prvých 10 000 filmov, avšak výsledky boli podobné aj pri tvorbe zhlukov, akurát sa do pôvodne väčších zhlukov dostalo viac filmov. Ďalej sme upustili od `directors, writers`, pretože nemali veľký vplyv vo výsledkoch PCA (keďže režisérov je veľa a transformácia do 0,1 stĺpcov vytvorí "riedku maticu", čo vo výsledku len spomaľovalo výpočet než dávalo lepší výstup)

3.1 PCA

Keďže veľa z našich atribútov je textových, navyše veľa z nich multikategorických (čiže viac kategórií v jednom zázname oddelených čiarkou), tak bolo treba pre lepšie zhlukovanie transformovať ich do tzv. *dummy variables*, a teda vytvoriť stĺpce s hodnotami 0, 1, prináležiac tomu, či záznam je z takejto kategórie alebo nie. Taktiež keby sme chceli analyzovať slová v názvoch filmov, museli by sme pridať ďalšie stĺpce podľa toho, či názov obsahuje dané slovo. To však výrazne zvýši rozmery našej tabuľky, a teda by bolo výhodnejšie ich zredukovať aby sa dalo lepšie rátať následné zhlukovanie. Na to sme využili PCA.

3.1.1 TF-IDF a najvýznamnejšie slová

Použitie všetkých slov z `primaryTitle` by príliš zvýšilo rozmery a taktiež veľa slov je bezvýznamných pre väčšinu filmov. Lepšie sa javí zvoliť podmožinu najdôležitejších slov pre filmy. CountVectorizer by mohol byť dobrou voľbou ale lepšiu informáciu o slovách vie zachytiť práve TF-IDF vectorizer a preto sme zvolili ten. Z najlepších slov sme odfiltrovali slová typu 'on', 'the' atď. a dostávame nasledovnú množinu slov:

night, love, house, black, part, life, with, star, christmas, movie, game, america, time, perfect, dead, from, evil, world, girls

Takéto dáta sme znormalizovali a spustili PCA, ktoré zachytáva minimálne 80% variácie v našich dátach - to vrátilo 33 komponentov. Ako najdôležitejšie vlastnosti sa ukazovali tieto (zoradené podľa

váhy):

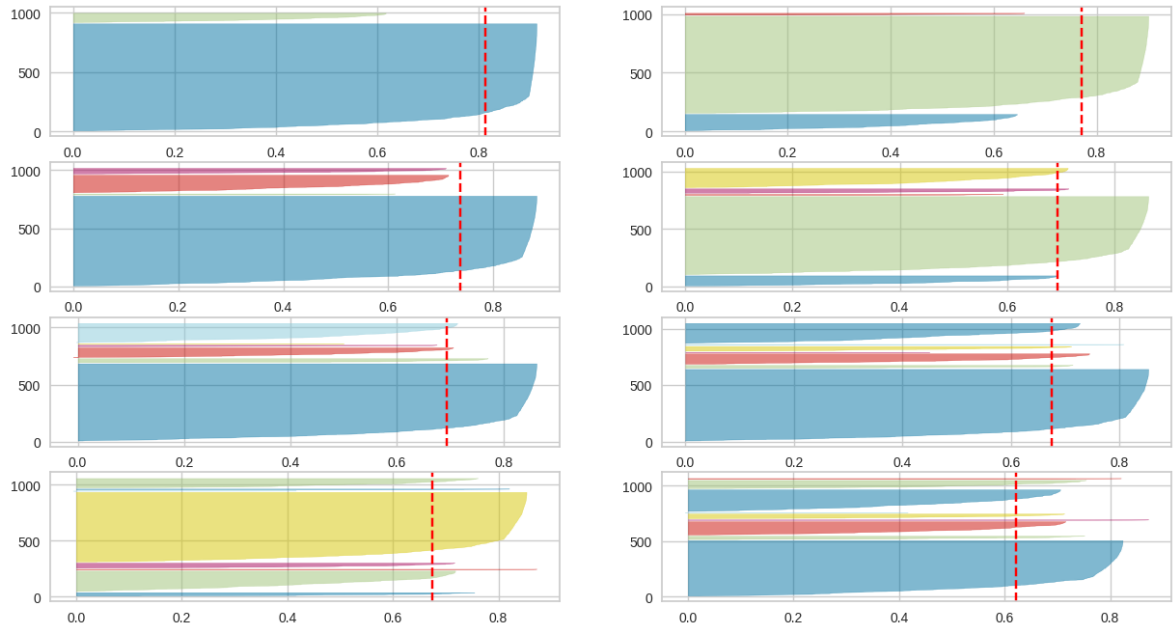
feature	absolute weight
type:movie	0.459178
genre:Adventure	0.366944
genre:Animation	0.365142
type:tvEpisode	0.323446
genre:Action	0.303529
averageRating	0.277332
type:tvSeries	0.262228
genre:Romance	0.182392
startYear	0.166592
genre:Drama	0.156904
type:video	0.121303
type:tvMiniSeries	0.095723
genre:Mystery	0.095519
genre:Horror	0.094765
genre:Comedy	0.089248
genre:History	0.070814
genre:Biography	0.068107
night	0.065056
genre:Musical	0.060215
type:short	0.053973

Tabuľka 3.1: Komponenty s najvyššou váhou pre popis dát

Je vidieť, že veľkú časť tvoria žánre, a teda žáner sa javí byť dôležitým faktorom spoločne aj s typmi. Zo slov vidíme v top len "night", takže slová vo všeobecnosti v názve filmov sa naopak nejavia byť až tak vplyvné.

3.2 Zhlukovanie

Na zhlukovanie sme použili obyčajný KMeans algoritmus. Na voľbu optimálneho k nezafungovala *elbow method* a teda sme použili tzv. *Silhouette method*, pri ktorej sa vykreslí graf s veľkosťami zhlukov a skóre a na základe toho zvolíme optimálne k.



Obr. 3.1: Silhouette graf pre $k \in [2, 10)$ ($k = 2$ v ľavom hornom, $k = 9$ v pravom dolnom rohu)

Zdá sa, že pri každom počte zhlukov nám vzniká jeden veľký zhluk a teda sa sústreďujeme najmä na dobré skóre a ako-takú rozmanitosť veľkosti zhlukov. To nás vedie k voľbe medzi 5 a 7 a keďže 5 má väčšiu rozmanitosť a podobnú silhouette skóre tak volíme 5. Vyšší počet nám už dáva nižšie silhouette skóre.

Dostávame rozdelenie do 5-tich zhlukov s takýmto priemerným hodnotením

zhluk	priemerné hodnotenie	počet filmov
3	8.200000	9
1	7.563158	38
4	7.082222	90
0	6.835119	672
2	6.788953	172

Tabuľka 3.2: Zhluky s ich priemerným hodnotením a početnosťou

Pričom priemerné hodnotenie všetkých riadkov je 6.8. Zoberieme teda najlepšie 3 zhluky (vyššie ako celkový priemer) a pozrieme sa na ich filmy a parametre keďže neobsahujú tak veľký počet filmov.

3.3 Výsledky

V zhluku 3 s najlepším hodnotením majú všetky filmy až na jednu výnimku hodnotenie > 8.0. Ich počet hodnotení je tiež najvyšší spomedzi všetkých zhlukov. Zdá sa, že KMeans nám minimálne vyselektovala nejakú podmožinu relatívne úspešných filmov. Všetky tieto filmy sú typu `movie` a ich

`runtimeMinute` je > 100 (medián 147). Väčšina bola nakrúcana po roku 2010 (s výnimkou The Big Lebowski(1998), The Good, the Bad and the Ugly(1966) a Kill Bill: Vol. 1(2003)). Avšak čo sa týka režisérov a spisovateľov (ktorí neboli zahrnutí v clusteringu kvôli zložitosti výpočtu) tak sa nezdajú byť žiadni spoloční pre viaceré filmy.

Pri ostatných zhlukoch je sa do `titleType` pridáva aj `tvSeries`, aj keď `movie` stále výrazne prevláda.

Pozrime sa na top žánre pre jednotlivé zhluky:

zhluk3	zhluk1	zhluk4
Action(5)	Action(20)	Drama(40)
Adventure(3)	Adventure(16)	Comedy(30)
Comedy(3)	Drama(16)	Action(28)
Crime(3)	Sci-Fi(10)	Adventure(26)
Drama(2)	Comedy(8)	Sci-Fi(16)

Tabuľka 3.3: Top žánre v zhlukoch a početnosť ich výskytov

Z tabuľky vidíme, že vo všetkých zhlukoch sa nachádzajú približne podobné žánre a top 3 by sa dali označiť `Action, Adventure, Comedy/Drama` (v predošlých analýzach je práve dvojica `Comedy` a `Drame` najúspešnejšia). Zhluky, ktoré obsahovali relatívne dobre hodnotené filmy mali v najlepších kategóriách obsadené vyššie spomínané žánre, akurát s rozdielom, že medzi najpočetnejšími sa vyskytovala `Action`, `Adventure` a až po nich nasledovali `Drama` a `Comedy`. Mohlo by sa zdať, že to spôsobil výber podmnožiny filmov, a teda že dané žánre neboli dostatočne reprezentované vo výbere, no v 1000 či 10 000 prvkovej podmnožine sú najpočetnejšie žánre práve `Drama` a `Comedy` a napriek tomu lepšie zhluky uprednostňujú `Action`, `Adventure`.

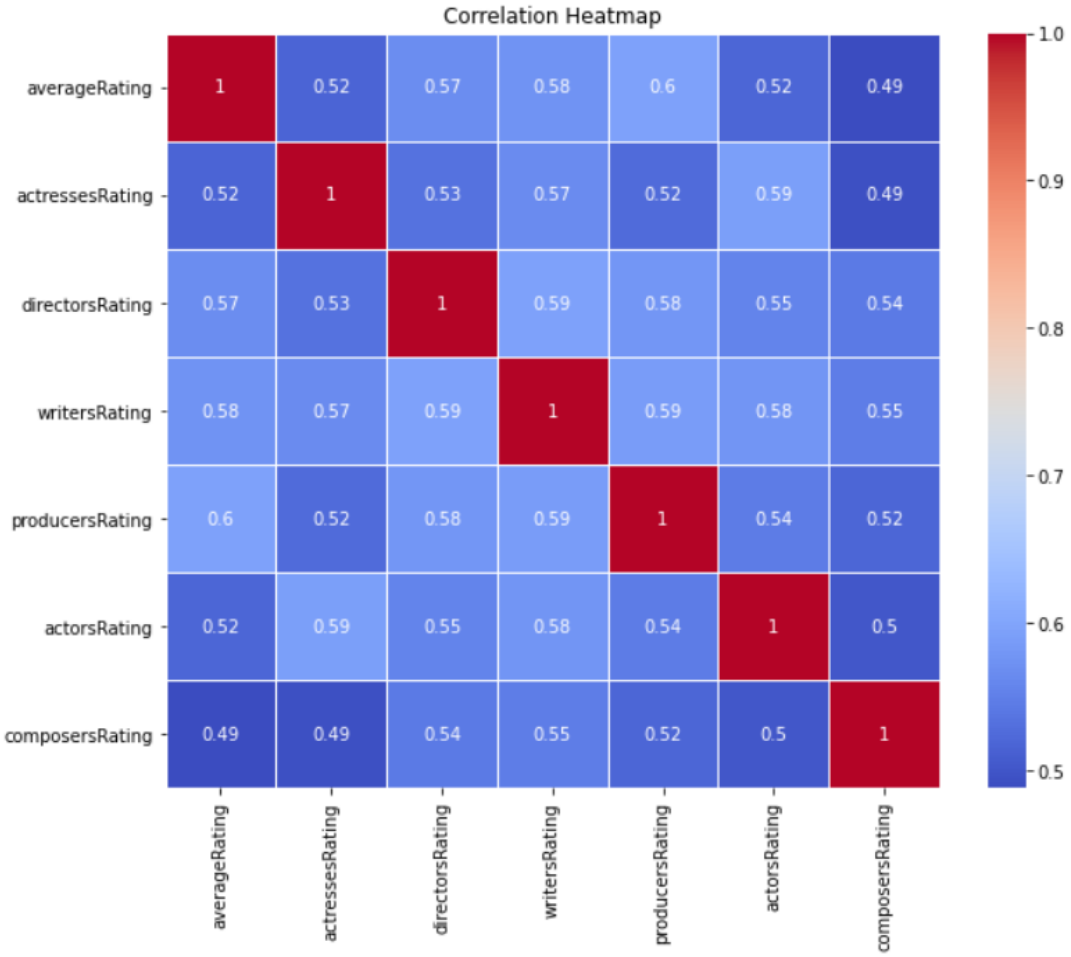
Zo zhlukov 1 a 4 sa variácia rokov vydania zvyšuje, avšak ich medián taktiež ostáva niekde blízko roku 2010. Teda kvalita filmov sa javí lepšia najmä pri produkcii v 21. storočí, avšak to nemusí znamenať, že rovnaký trend nemajú aj zlé filmy (bolo by potrebné podobnú analýzu spraviť aj na filmoch s menším ohlasom). Iba to potvrdzuje aj naše tušenie z predošlých analýz. Taktiež medián dĺžky filmu v oboch je vyšší než 100, aj keď sa v zhlukoch už objavujú aj kratšie filmy.

Čo sa týka režisérov, tí sa zdajú byť disjunktní v zhlukoch avšak v zhlukoch 1 a 4 (pravdepodobne najmä kvôli ich vyššej početnosti) sú aj režiséri ktorí sa objavujú 3 krát (Steven Spielberg pre zhluk 1; John Dahl a Tim Burton pre zhluk 4). V zhluku 3 s najvyšším priemerným hodnotením sú top 5 títo režiséri: Sergio Leone, Bong Joon Ho, Martin Scorsese, Quentin Tarantino a Jon Watts.

Kapitola 4

Vplyv ľudí na obľúbenosť filmov

V tejto časti sme sa pozreli aký vplyv majú jednotliví ľudia pre filmy. Táto otázka je celkom zložitá, pretože ako vieme posúdiť kvalitu herca/režiséra/niekoho iného alebo v našom prípade kvalitu viacerých hercov spoločne hrajúcich v jednom filme? Kvalitu herca sme vypočítali tak, že sme sa pozreli na všetky filmy, v ktorých daný človek hral a vyrátali priemer ich hodnotení. To isté sme spravili pre režisérov, producentov, scenáristov, herečky a skladateľov. Teraz ak vo filme hralo viac hercov, tak sme kvalitu hercov daného filmu spočítali ako priemer kvalít jednotlivých hercov. To isté aj pre ostatné roly. Toto by nám ale mohlo dávať dosť skreslené výsledky pri výpočte kvality hercov v danom filme, keďže do ich kvality by bola zahrnutá aj kvalita daného filmu, preto tú sme pri ich výpočte nebrali do úvahy. Po tomto sme sa pozreli, že kvalita ktorej roly najviac vplývala na kvalitu filmu. Tu nám najvyššia korelácia vyšla pri producentovi, až 0.64.1. Najmenej vplývala kvalita skladateľa. Tiež tu môžeme vidieť zaujímavú vec, že kvalita herečiek mala s hodnotením filmu rovnakú koreláciu ako kvalita hercov. Táto metodika by sa určite dala zlepšiť, keby sme brali dáta iba z lepších filmov (s najväčšími rozpočtami/najväčším počtom hlasov). Ďalej pri výpočte kvality hereckého obsadenia majú všetci herci rovnakú váhu, čo nie je úplne ideálne, keďže niektorí herci sa vo filme vyskytujú viac. Radšej by sme mohli urobiť vážený priemer, lenže sme nemali údaj o tom, ktorý herec sa ako dlho vyskytoval v danom filme.



Obr. 4.1: Heatmapa korelácií hodnotení

Kapitola 5

Záver

V prvej časti sme skúmali vplyv žánrov na obľúbenosť filmov. Počet filmov pre jednotlivé žánre sa nezdá byť ako vplyvný faktor. Následne sme pomocou regresného modelu odhadli, ktoré žánre by mohli byť obľúbené a neobľúbené. Najviac obľúbené žánre sú `Drama`, `Comedy` a najmenej obľúbené `Adult`, `Thriller` a `Horror`. Z týchto výsledkov sme usúdili, že žáner má vplyv na obľúbenosť.

Ďalším podkladom pre toto tvrdenie by bolo, že aj pri PCA nám vyšla väčšina žánrov ako dôležité komponenty. V zhlukovaní tieto žánre vyšli tiež dôležité, avšak pred nimi sa javili ako úspešnejšie `Action` a `Adventure`.

Ďalej sme zistili, že filmy typu `movie` majú v priemere nižšie hodnotenie a `tvEpisode` vyššie.

Pri zhlukovaní sme si taktiež všimli, že úspešnejšie sú najmä filmy s `Runtime` väčším ako 100 (niekde okolo 130). Taktiež typ `movie` výrazne prevláda pred ostatnými formátmi ako `tvSeries` alebo `short`.

Rok vydania filmu sa ukázal ako vplyvný faktor u lepšie hodnotených žánrov, a preto aj tento faktor považujeme za významný pre obľúbenosť filmov.

Vplyv hercov, režisérov, skladateľov a iných sa ukázal tiež ako celkom ovplyvňujúci faktor na hodnotenie, avšak nie úplne smerodajným, keďže sme nemali dostupné hodnotenia jednotlivých hercov ale sme si ich museli vyrobiť na základe hodnotenia filmov.

Z našich výsledkov je možné usúdiť, že niektoré faktory ako žáner či typ vedia ovplyvniť celkové hodnotenie filmu.