Krzysztof Szafrański
Nr indeksu: 386 118

Determinanty wysokości przychodów filmów

Celem niniejszego projektu jest pokazanie jakie zmienne wpływają na kasowy sukces filmów w kinach. Analizę przeprowadzono na zbiorze danych zawierającym informacje na temat ponad 7000 filmów z lat 1980-2020, który pochodzi ze strony: https://www.kaggle.com/datasets/danielgrijalvas/movies. Poniżej przedstawiono krótki opis analizowanych zmiennych:

- *gross* - przychody jakie dany film osiągnął w amerykańskich kinach
- *rating* - kategoria wiekowa
- *genre* - gatunek filmu
- *year* - rok produkcji
- *score* - ocena filmu na stronie IMDb
- *votes* - liczba ocen na IMDb
- *country* - kraj produkcji
- *budget* - budżet filmu
- *company* - producent filmu
- *runtime* - długość filmu

Wstępna analiza danych

W pierwszej kolejności przeanalizowano dla jakich poziomów zmiennych kategorycznych średnie przychody filmów są największe. W przypadku zmiennej *rating* średnie przychody są większe dla produkcji o niższych ograniczeniach wiekowych. Nie jest to zaskakujące, ponieważ wyższe kategorie wiekowe siłą rzeczy ograniczają liczbę widzów.

	gross	
	count	mean
rating		
G	153	1.413539e+08
PG-13	2092	1.308173e+08
PG	1226	1.065848e+08
R	3613	4.266882e+07
NC-17	35	2.815856e+07
Unrated	306	1.530908e+07

Natomiast najlepiej sprzedające się gatunki, to: animacje, filmy familijne, filmy akcji oraz filmy przygodowe. Warto jednak zwrócić uwagę, że dla niektórych gatunków jest bardzo mało obserwacji.

	gross	
	count	mean
genre		
Animation	335	2.392300e+08
Family	11	1.961725e+08
Action	1673	1.455086e+08
Adventure	420	1.093252e+08
Mystery	20	1.011835e+08
Biography	433	4.787432e+07
Horror	307	4.737241e+07
Comedy	2192	4.433187e+07
Crime	542	3.940120e+07
Drama	1468	3.893096e+07
Fantasy	43	3.870933e+07
Sci-Fi	8	3.256123e+07
Thriller	12	2.693526e+07
Romance	8	2.354937e+07
Western	3	1.067530e+07
Musical	2	2.595346e+06
Sport	1	1.067629e+06
Music	1	1.100140e+05
History	0	NaN

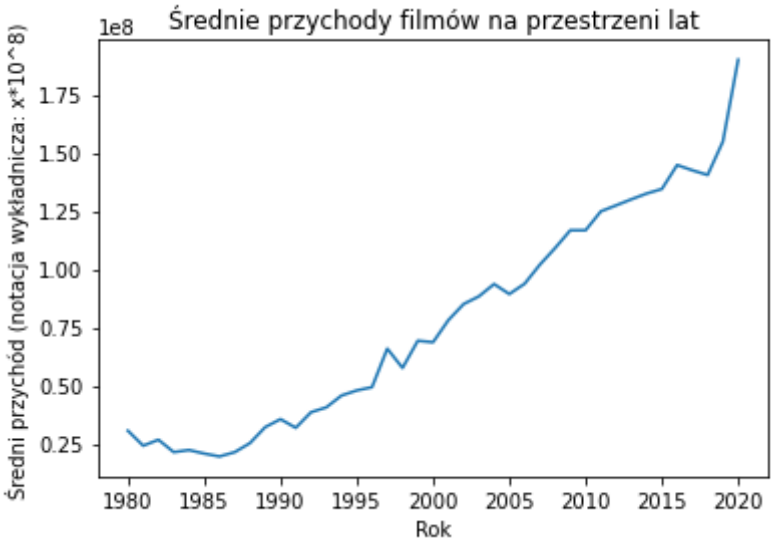
Poniższa tabela zawiera 10 wytwórni filmowych, których filmy osiągają największe średnie przychody. Jednak ponownie dla części wytwórni jest tylko jedna obserwacja.

	gross	
	count	mean
company		
Marvel Studios	12	1.255466e+09
Illumination Entertainment	2	1.097122e+09
Fairview Entertainment	1	9.665549e+08
B24	1	8.806815e+08
Avi Arad Productions	1	8.560852e+08
Chris Morgan Productions	1	7.590569e+08
Jolie Pas	1	7.584118e+08
Coco Cartoon	1	7.262641e+08
Lucasfilm	10	7.185352e+08
Marvel Entertainment	1	7.144215e+08

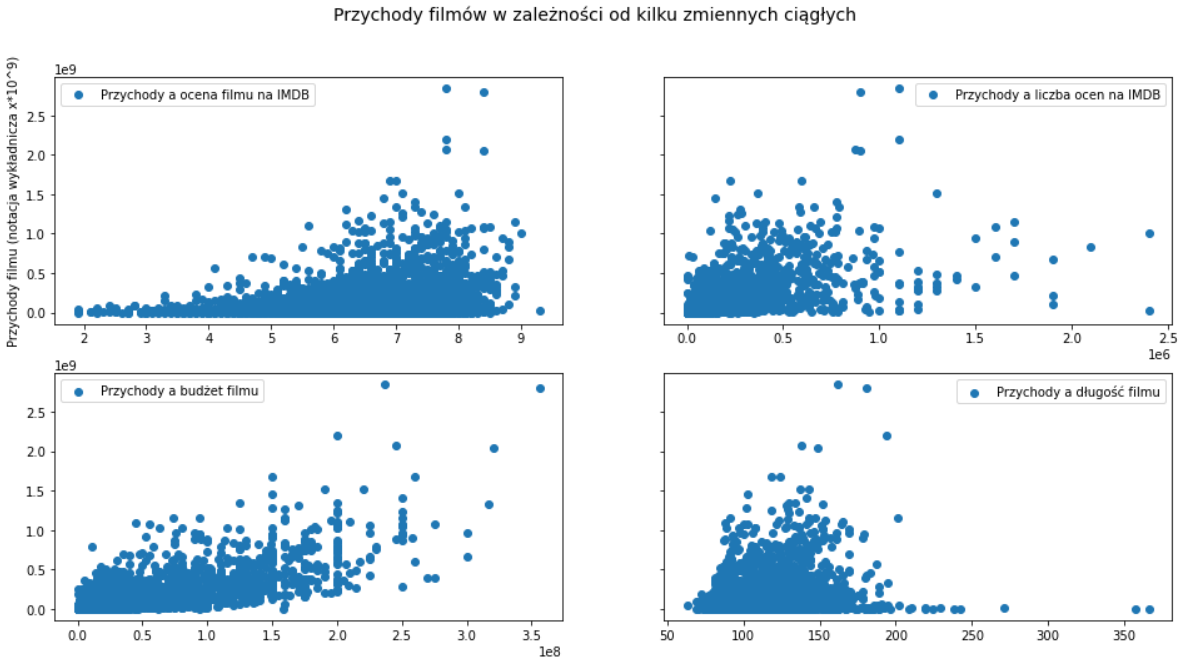
Poniższej przedstawiono państwa, gdzie produkowane filmy osiągały największe średnie przychody. Co ciekawe, Stany Zjednoczone są pod tym względem dopiero na 5 miejscu. Wyprzedzają je m.in. Nowa Zelandia i Chiny.

	gross	
	count	mean
country		
Malta	1	3.527941e+08
New Zealand	24	2.647805e+08
China	40	2.197357e+08
Finland	3	1.691938e+08
United States	5377	8.991236e+07
United Arab Emirates	2	8.858613e+07
South Africa	7	8.102684e+07
Lebanon	1	6.441700e+07
United Kingdom	798	6.134110e+07
Germany	116	5.360854e+07

W następnej kolejności zajęto się analizą zmiennych ciągłych. Od drugiej połowy lat 80 średnie przychody filmów niemal nieprzerwanie rosną, co zostało przedstawione na poniższym wykresie.

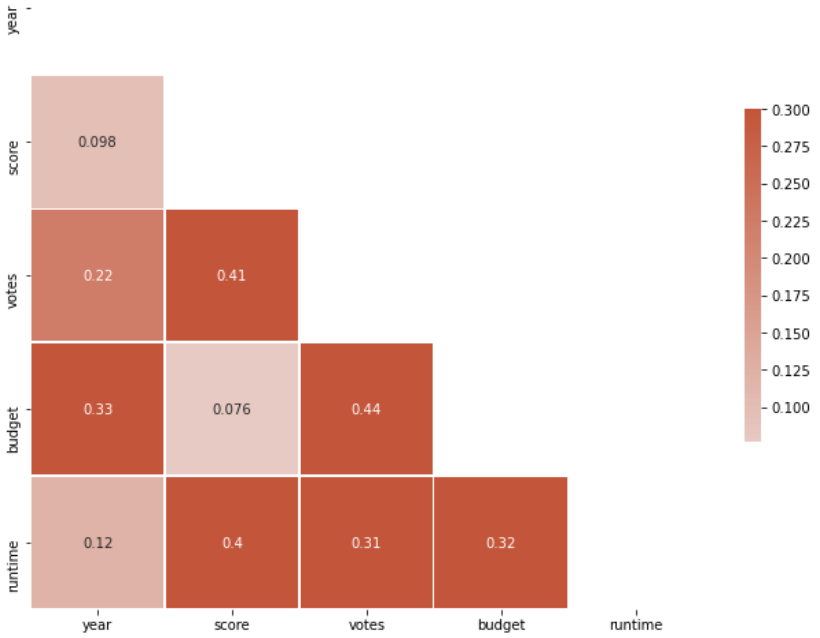


Poniżej przedstawiono wykresy punktowe, pokazujące relacje między pozostałymi zmiennymi a przychodami filmów. Wydaje się, że zachodzi pozytywna korelacja między przychodami a oceną filmu oraz jego budżetem.

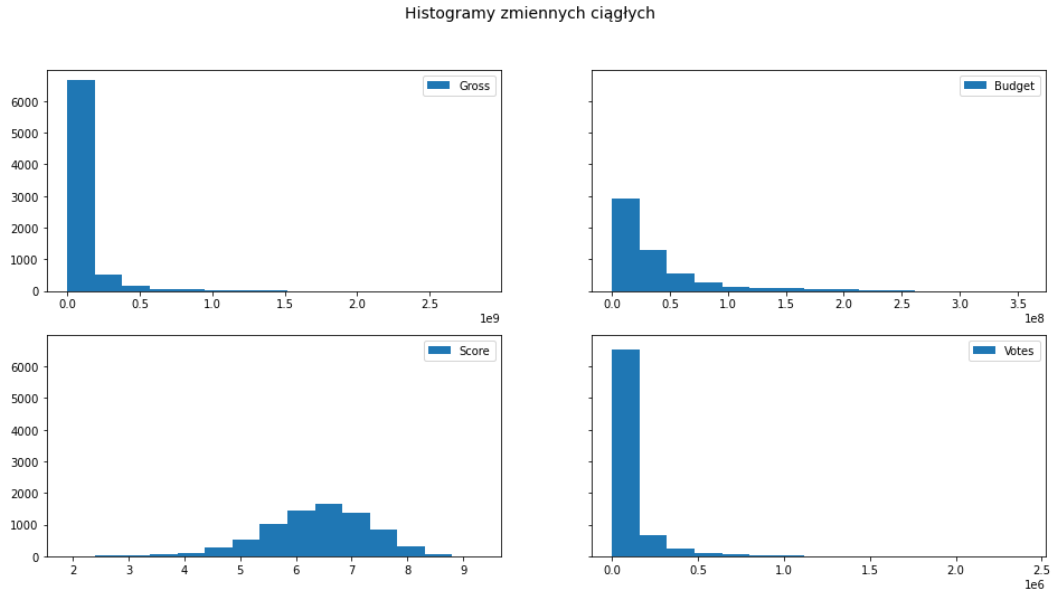


Model ekonometryczny

Zdecydowano się również przeprowadzić prostą regresję liniową. W modelu nie umieszczono zmiennych *company* i *country*, ze względu na bardzo dużą liczbę poziomów tych zmiennych. Nie ma dużej korelacji pomiędzy zmiennymi ciągłymi, więc wydaje się, że współliniowość nie powinna być istotnym problemem.



Poniższe histogramy sugerują, że rozkład niektórych zmiennych zdecydowanie odbiega od normalnego. Zdecydowano się więc zlogarytmować zmienną zależną *gross* oraz zmienne objaśniające *budget* i *votes*.



Poniżej przedstawiono wyniki przeprowadzonej regresji metodą najmniejszych kwadratów. Statystyka F wskazuje, że wszystkie zmienne są łącznie istotnie. Model wyjaśnia 67% całkowitej zmienności przychodów filmów.

OLS Regression Results

Dep. Variable:	np.log(gross)	R-squared:	0.679
Model:	OLS	Adj. R-squared:	0.678
Method:	Least Squares	F-statistic:	476.5
Date:	Mon, 20 Jun 2022	Prob (F-statistic):	0.00
Time:	12:47:34	Log-Likelihood:	-8055.6
No. Observations:	5423	AIC:	1.616e+04
Df Residuals:	5398	BIC:	1.633e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	17.8922	3.251	5.504	0.000	11.519	24.265
C(rating)[T.NC-17]	-1.3572	0.303	-4.483	0.000	-1.951	-0.764
C(rating)[T.PG]	-0.0971	0.116	-0.836	0.403	-0.325	0.131
C(rating)[T.PG-13]	-0.3535	0.123	-2.883	0.004	-0.594	-0.113
C(rating)[T.R]	-0.7488	0.122	-6.139	0.000	-0.988	-0.510
C(rating)[T.Unrated]	-1.6919	0.186	-9.082	0.000	-2.057	-1.327
C(genre)[T.Adventure]	-0.1330	0.069	-1.926	0.054	-0.268	0.002
C(genre)[T.Animation]	0.3928	0.090	4.355	0.000	0.216	0.570
C(genre)[T.Biography]	-0.1528	0.071	-2.144	0.032	-0.293	-0.013
C(genre)[T.Comedy]	0.0462	0.042	1.097	0.273	-0.036	0.129
C(genre)[T.Crime]	-0.1880	0.063	-2.966	0.003	-0.312	-0.064
C(genre)[T.Drama]	-0.1303	0.050	-2.616	0.009	-0.228	-0.033
C(genre)[T.Family]	0.3257	0.539	0.605	0.545	-0.730	1.382
C(genre)[T.Fantasy]	-0.0722	0.171	-0.423	0.672	-0.407	0.262
C(genre)[T.History]	3.366e-13	2.1e-13	1.601	0.110	-7.56e-14	7.49e-13
C(genre)[T.Horror]	0.3507	0.077	4.542	0.000	0.199	0.502
C(genre)[T.Music]	2.452e-13	1.53e-13	1.598	0.110	-5.55e-14	5.46e-13
C(genre)[T.Musical]	-2.25e-13	1.41e-13	-1.599	0.110	-5.01e-13	5.08e-14
C(genre)[T.Mystery]	-0.3540	0.262	-1.353	0.176	-0.867	0.159
C(genre)[T.Romance]	-1.1770	0.481	-2.447	0.014	-2.120	-0.234
C(genre)[T.Sci-Fi]	0.0191	0.439	0.043	0.965	-0.841	0.879
C(genre)[T.Sport]	-2.795e-15	1.74e-15	-1.608	0.108	-6.2e-15	6.13e-16
C(genre)[T.Thriller]	0.6915	0.406	1.701	0.089	-0.105	1.488
C(genre)[T.Western]	0.6590	0.759	0.868	0.385	-0.829	2.147
year	-0.0081	0.002	-4.899	0.000	-0.011	-0.005
score	-0.1585	0.022	-7.298	0.000	-0.201	-0.116
np.log(votes)	0.7731	0.015	50.333	0.000	0.743	0.803
np.log(budget)	0.4984	0.016	30.252	0.000	0.466	0.531
runtime	0.0034	0.001	3.225	0.001	0.001	0.006

Omnibus:	1602.193	Durbin-Watson:	1.890
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9150.161
Skew:	-1.291	Prob(JB):	0.00
Kurtosis:	8.816	Cond. No.	1.00e+16

Prawie wszystkie kategorie wiekowe są powiązane z mniejszymi przychodami niż kategoria G, która jest poziomem bazowym. Jedynie kategoria PG okazała się być nieistotna. W przypadku gatunku poziomem bazowym są filmy akcji. Jedynymi gatunkami, które

osiągają średnio większe przychody są animacje oraz horrory. Pozostałe gatunki wiążą się z niższymi przychodami bądź okazały się nieistotne.

Co ciekawe, wzrost roku produkcji o 1 jest powiązany ze spadkiem przychodów o niecały 1%. Podobnie wraz ze wzrostem oceny filmu o 1 przychody maleją o 15%. Prowadzi to do odmiennych wniosków niż tych wynikających z przeprowadzonej wcześniej wstępnej analizy danych.

Natomiast wzrost liczby ocen o 1% jest powiązany ze wzrostem przychodów o 0,77% a wzrost budżetu filmu o 1% również wiąże się ze wzrostem przychodów o 0,5%. Ponadto wraz ze wzrostem długości filmu o minutę przychody rosną o niespełna 0,5%.

Podsumowanie

Przeprowadzona analiza sugeruje, że z przychodami osiąganymi przez filmy pozytywnie związany jest budżet filmu oraz liczba ocen filmu na IMDb. Ponadto na większy sukces kasowy mogą liczyć filmy bez ograniczenia wiekowego. Natomiast najpopularniejsze gatunki to: animacje, filmy akcji i horrory. Zdecydowanie mniejszym zainteresowaniem widzów cieszą się kryminały i dramaty.