Statistična analiza filmov

Jure Bevc

januar 2018

1 Uvod

Dandanes se v filmski industriji snemajo filmi, ki za svojo izdelavo zahtevajo več milijonov dolarjev sredstev. Zelo priročno bi bilo, če bi lahko v naprej določili uspešnost filma in ugotovili, koliko dobička lahko zanj pričakujemo. S pomočjo statistične analize in zbirke podatkov, ki vsebuje podatke več tisočih filmov, lahko poskušamo ugotoviti, ali se med podatki pojavijo kakšni vzorci oziroma odvisnosti.
Cilj te naloge je namreč uporabiti znanje verjetnosti in statistike na resničnih podatkih in ne na izmišljenih ali računalniško generiranih.

1.1 Opredelitev problema

Ugotoviti želimo, kako se je proračun filmov spreminjal skozi čas, kako so porazdeljene ocene filmov in kaj lahko sklepamo o dobičku. Zanima nas namreč vplivnost posameznih lastnosti filmov na dobiček. Tako lahko ugotovimo, da imajo nekatere lastnosti pomembnejšo vlogo pri izdelavi.

1.2 Pridobivanje podatkov in orodja

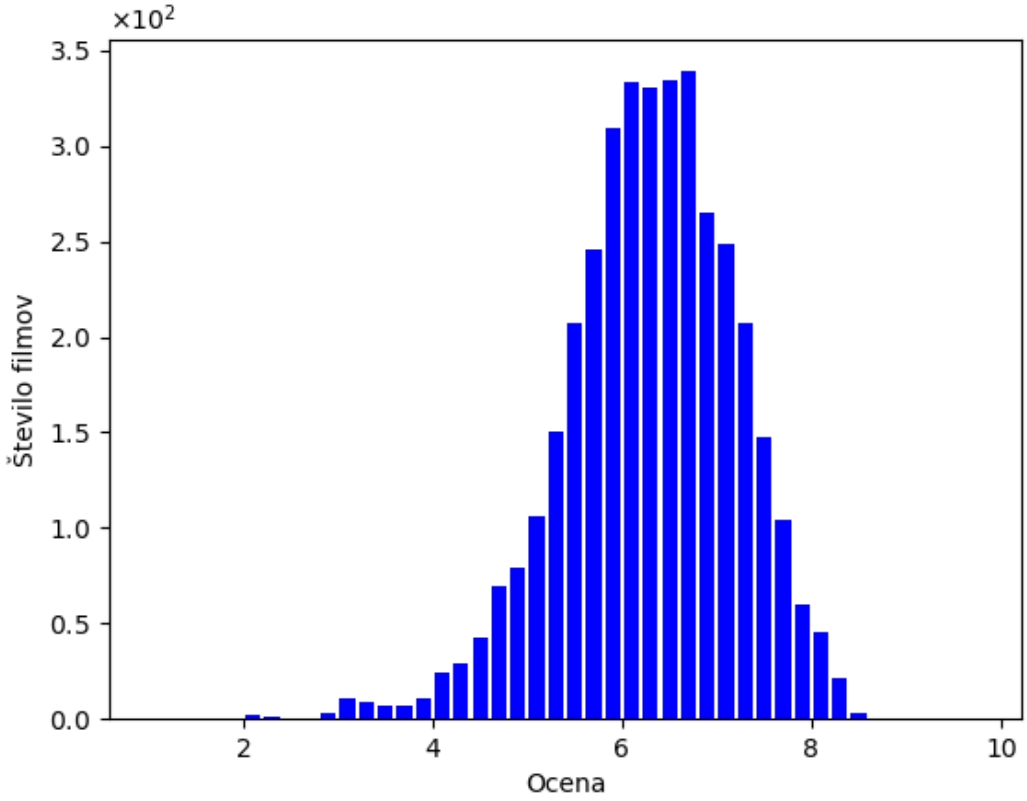
Za smiselno analizo potrebujemo čim več filmov in njihove podatke. Na spletni strani *Kaggle*[2] je že pripravljena priročna podatkovna zbirka, ki vsebuje približno 5000 filmov[4]. V zbirki podatkov se pojavijo tudi filmi, ki nimajo vseh podatkov (manjkajoče ocene, proračuni, itd.), ki jih potrebujemo za analizo, zato je končno število našega vzorca okoli 3700 filmov.
Vsi proračuni in prihodki filmov ter izračuni vrednosti v tej nalogi so merjeni v dolarjih.

Za obdelavo takšne količine podatkov lahko uporabimo programski jezik *Python* in knjižnici *Matplotlib* ter *NumPy*. S tem si olajšamo postopek risanja grafov in skrajšamo nekatere računske dele analize.

2 Analiza

2.1 Distribucija ocen

Ko se odločamo za ogled filma, je pogosto pomembna njegova ocena. Oceno filmov podajo gledalci, ki so si film že ogledali. V našem primeru so to povprečne ocene uporabnikov spletne strani TMDb[4], ki imajo vrednosti med 0 in 10.
V spodnjem grafu so prikazane ocene in število filmov, ki so dosegli takšno vrednost ocene.



Slika 1: Distribucija ocen

Graf ima obliko podobno binomski porazdelitvi, zato jo lahko primerjamo z naravno porazdelitvijo in lahko izračunamo pričakovano vrednost ter standardni odklon tega vzorca.
Pričakovano vrednost dobimo po spodnji formuli, kjer je v našem primeru n število ocen in x_i število filmov s to oceno:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 6,228$$

Povprečna ocena filmov v našem vzorcu je torej $6,228$. Popravljeni vzorčni standardni odklon[1] pa dobimo po formuli:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = 0,923$$

2.1.1 Interval zaupanja

Z izračunano pričakovano vrednostjo in standardnim odklonom lahko določimo še interval zaupanja za vrednost μ celotne populacije filmov. Ker natančen standardni odklon σ populacije ni znan, uporabimo spodnjo formulo s Studentovo porazdelitvijo[1], za stopnjo zaupanja β pa izberemo 95%:

$$\Delta = \frac{s \cdot t_{(1+\beta)/2}(n-1)}{\sqrt{n}} = \frac{0,923 \cdot t_{1,95/2}(3764-1)}{\sqrt{3764}} = 0,0295$$

Interval zaupanja s stopnjo $\beta = 95\%$ je torej:

$$\overline{x} - \Delta \leq \mu \leq \overline{x} + \Delta$$
$$6,228 - 0,0295 \leq \mu \leq 6,228 + 0,0295$$
$$6,1985 \leq \mu \leq 6,2575$$

Prav tako pa lahko pri enaki stopnji zaupanja izračunamo interval za standarni odklon σ:

$$c_1 = \chi^2_{(1-\beta)/2} \cdot (n-1) = \chi^2_{0,025}(3763) = 3594,87$$

$$c_2 = \chi^2_{(1+\beta)/2} \cdot (n-1) = \chi^2_{0,975}(3763) = 3934,92$$

$$s \cdot \sqrt{\frac{n-1}{c_2}} \leq \sigma \leq s \cdot \sqrt{\frac{n-1}{c_1}}$$
$$0,923 \cdot \sqrt{\frac{3763}{3934,92}} \leq \sigma \leq 0,923 \cdot \sqrt{\frac{3763}{3594,87}}$$
$$0,90 \leq \sigma \leq 1,05$$

2.1.2 Testna statistika

Lahko bi sumili, da bo povprečna ocena vseh filmov okoli $5,0$, ker je ravno na sredini vrednosti možnih ocen. Domnevo, da je ocena vseh filmov ravno takšna, lahko preverimo s testno statistiko[1]. Na začetku postavimo našo ničelno domnevo H_0, ki trdi, da je povprečna ocena $5,0$, nato pa zapišemo še alternativno domnevo H_1, ki trdi, da je povprečna ocena v resnici večja:

$$H_0 : \mu = 5,0 \qquad H_1 : \mu > 5,0$$

Ker imamo precej velik vzorec, za območje zavračanja[1] vzamemo $\alpha = 0,01$. Nato izračunamo testno statistiko po formuli:

$$TS = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{6,228 - 5,0}{0,923/\sqrt{3764}} = 81,62$$

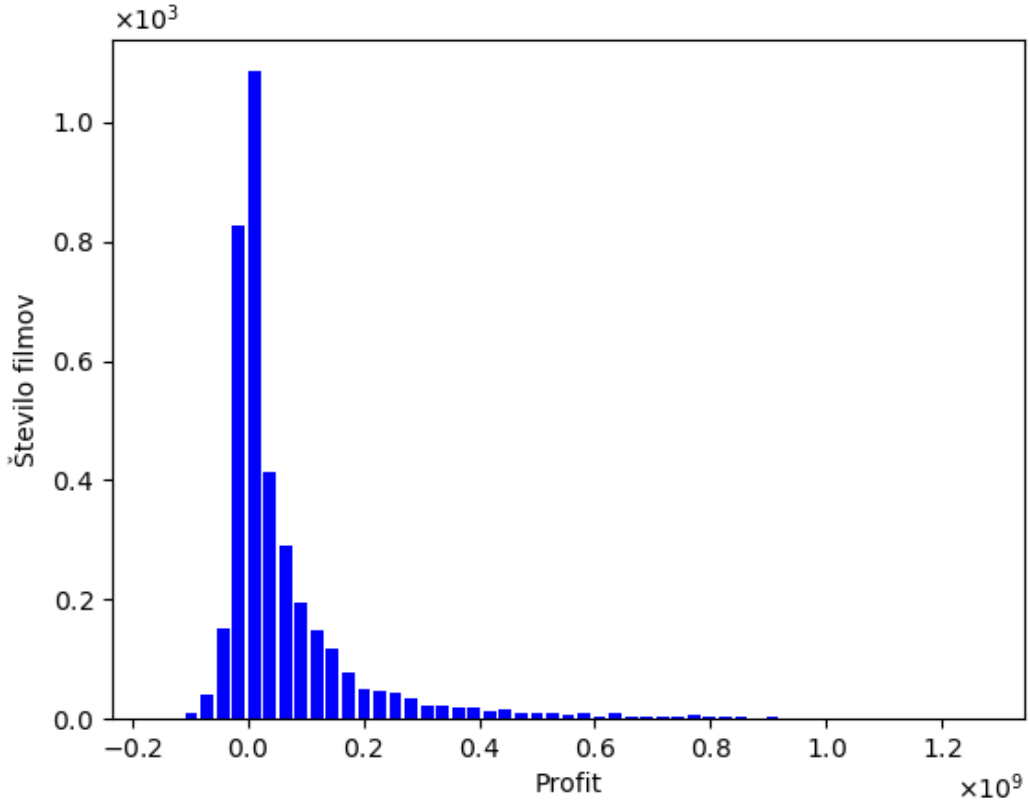
Pogledamo v tabelo z porazdelitve in izračunamo $P - vrednost$, ki pa bo v našem primeru izjemno majhna, saj je testna statistika prevelika, da bi sploh bila napisana v tabeli z porazdelitve:

$$P - vrednost << 0,00001$$

Ker velja $P - vrednost < \alpha$, lahko zavrnemo ničelno domnevo in sprejmemo alternativno. Povprečna ocena vseh filmov je torej gotovo nad $5,0$.

2.2 Dobiček filmov

Za podjetje, ki izdeluje filme, je pogosto pomemben dobiček, ki ga prinese izdelan film. Zato nas zanima, ali je dobiček morda povezan s kakšno drugo lastnostjo filma, na katero lahko izdelovalci filmov neposredno vplivajo. Dobiček filma izračunamo tako, da od celotnega prihodka odštejemo proračun filma. Oba podatka sta v naši zbirki podatkov[4].
Za začetek si oglejmo kako je dobiček vseh filmov porazdeljen.



Slika 2: Distribucija dobičkov

Graf ima strmo zvonasto obliko in en vrh, zato lahko ponovno izračunamo povprečje in standardni odklon kot v prejšnjem primeru. Tokrat je x_i profit posameznega filma, n pa število filmov v vzorcu:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 66\ 951\ 430$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} = 150\ 425\ 789$$

Povprečni dobiček filmov našega vzorca je torej približno 67 milijonov dolarjev, lahko pa še določimo interval zaupanja, za boljšo predstavo povprečja o celotni populaciji. Ponovno uporabimo stopnjo zaupanja $\beta = 95\%$.

$$\Delta = \frac{s \cdot t_{(1+\beta)/2}(n-1)}{\sqrt{n}} = \frac{150425789 \cdot t_{1,95/2}(3764-1)}{\sqrt{3764}} = 4\ 805\ 666$$

$$\overline{x} - \Delta \leq \mu \leq \overline{x} + \Delta$$

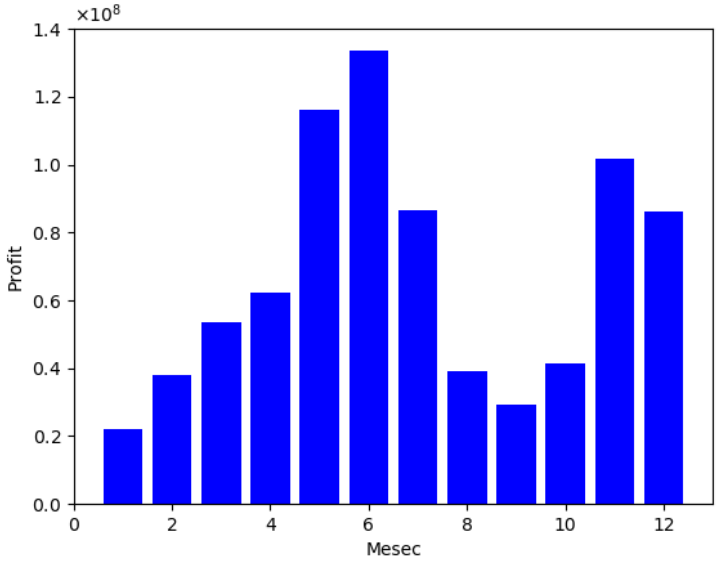
$$66\ 951\ 430 - 4\ 805\ 666 \leq \mu \leq 66\ 951\ 430 + 4\ 805\ 666$$

$$62\ 145\ 764 \leq \mu \leq 71\ 757\ 096$$

Povprečni dobiček vseh filmov je torej med $62,1$ in $71,7$ milijonov z verjetnostjo 95%

2.2.1 Dobiček in mesec izdaje

Ena od lastnosti, na katero lahko izdelovalci filmov običajno vplivajo, je čas izdaje filma. Ko je film v popolnosti izdelan, načeloma ne želimo čakati predolgo do izdaje, zato si bomo v tem primeru ogledali, ali je mesec izdaje pomemben pri dobičku. Spodnji graf prikazuje povprečni dobiček filmov, ločen po mesecih izdaje filma.

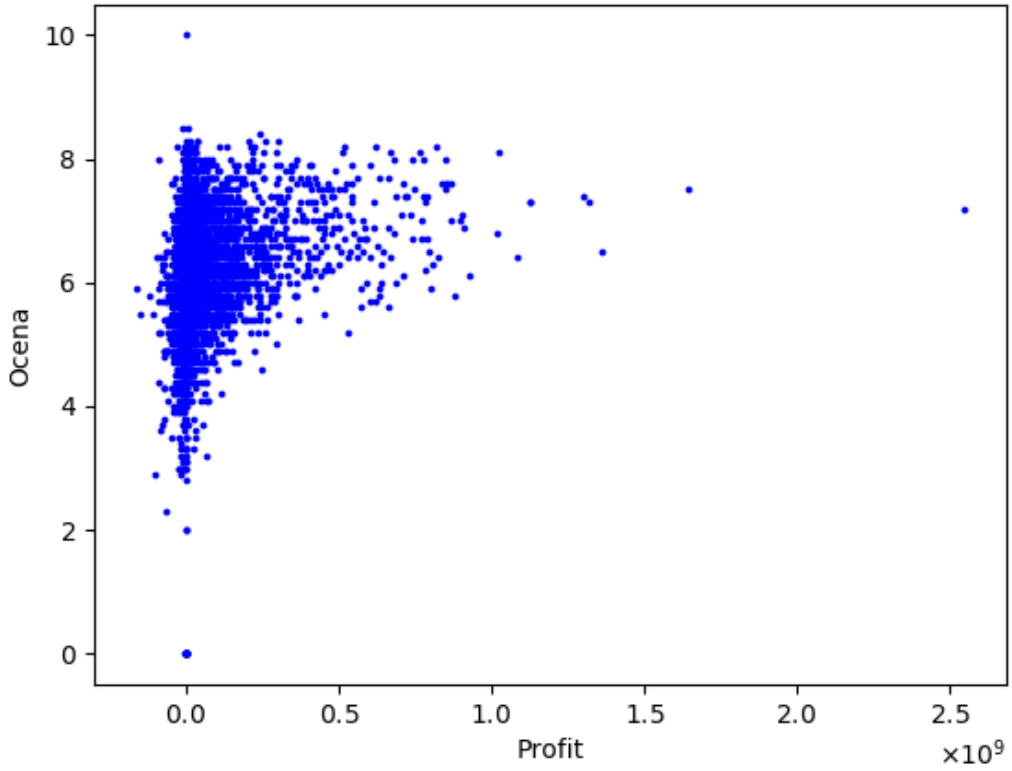


Slika 3: Povprečni dobički filmov po mesecih

Opazimo, da se dobički v posameznih mesecih med seboj precej razlikujejo. V našem vzorcu imajo večji povprečni dobiček tisti filmi, ki so bili izdani na sredini ali proti koncu leta.

2.2.2 Odvisnost dobička in ocene

V prejšnji točki smo ugotovili, da so za večji dobiček nekateri meseci primernejši od drugih. Lahko pa si ogledamo še, ali je ocena filma odvisna od dobička, ali pa je morda zanj popolnoma nepomembna. Naslednji graf prikazuje profit in oceno filmov našega vzorca.



Slika 4: Dobiček filma v odvisnosti od ocene, ki jo dobi

Ker želimo ugotoviti korelacijo med dobičkom in ceno, je smiselno za to uporabiti korelacijski koeficient r_{XY}. Ta se izračuna po spodnji formuli in nakazuje na močno korelacijo, ko ima vrednost blizu 1 ali -1, in zanemarljivo korelacijo, ko ima vrednost blizu 0:

$$r_{XY} = \frac{k(X,Y)}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}$$

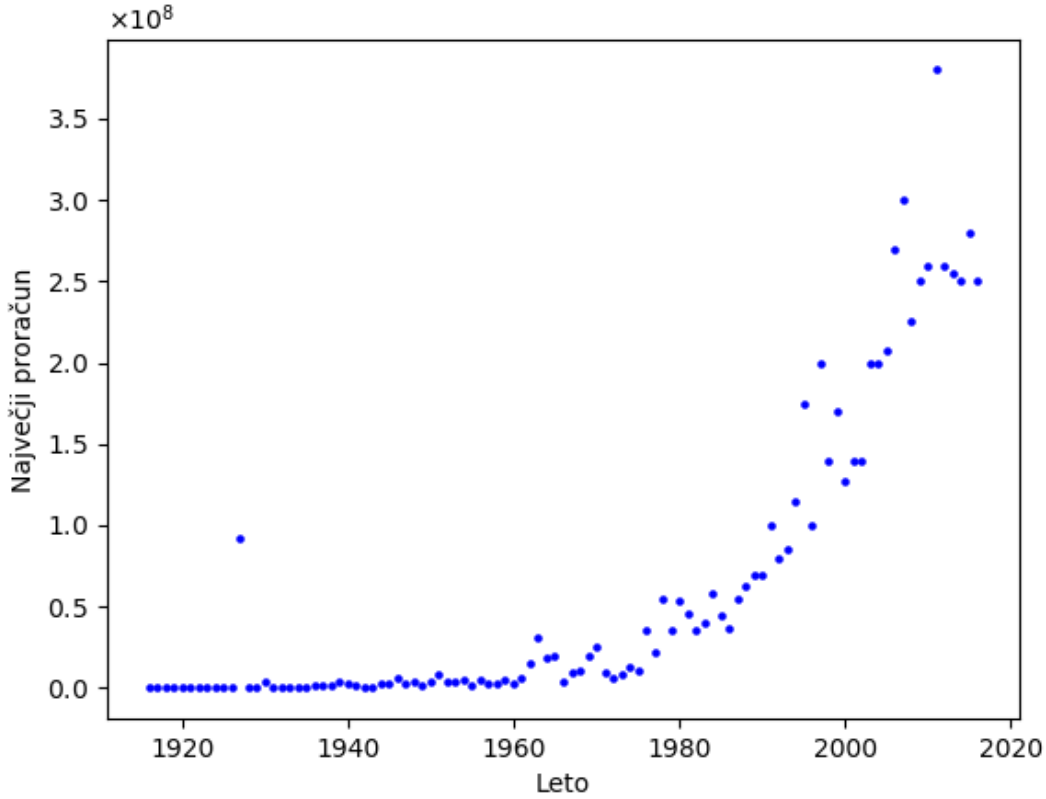
Ker je število podatkov veliko, lahko velike vsote, ki se pojavijo v formuli, izračunamo s pomočjo računalnika, da dobimo:

$$r_{XY} = \frac{127812199475}{8.51 \cdot 10^{19} \cdot 3475} = 4.31 \cdot 10^{-13}$$

Korelacijski koeficient je izjemno blizu vrednosti 0, zato ne moramo sklepati njune soodvisnosti. To pomeni, da ocena, ki jo dodelijo gledalci filmu, načeloma ni močno povezana s profitom, ki ga lahko pričakujemo od filma.

2.3 Proračun in leto izdaje filmov

Še en podatek, ki si ga je zanimivo ogledati, je proračun filmov in kako se ta spreminja skozi čas. Danes namreč vemo, da za izdelavo filma večja podjetja vložijo več milijonov dolarjev, zanima pa nas tudi to, kako je bilo v preteklosti in kaj lahko sklepamo o prihodnosti. Naslednji graf prikazuje največje proračune v določenem letu:



Slika 5: Največji proračun skozi leta

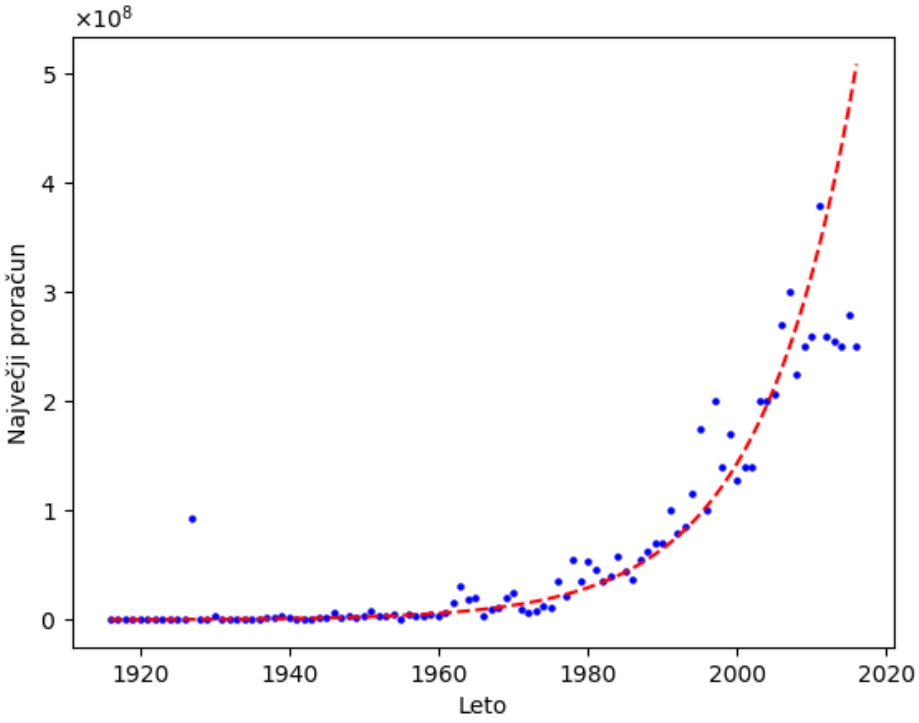
Opazimo, da maksimalni proračun iz leta v leto raste in da je njegov graf precej podoben eksponenti funkciji. S pomočjo linearne regresije lahko aproksimiramo eksponentno funkcijo oblike $c \cdot e^x$ na naš graf. S tem bomo lahko ocenili največji proračun, ki bo razpisan v letu 2018. Ker ima graf preveč točk, da bi funkcijo aproksimirali ročno, lahko to storimo s pomočjo knjižnice *Numpy*.

$$parametri = np.polyfit(leto, np.log(proracun), 1)$$

V tem primeru smo logaritmirali funkcijo $proracun = c \cdot e^{leto}$ in dobili $log(proracun) = log(c) \cdot leto$, nato pa funkcija $np.polyfit()$ poskrbi za iskanje parametrov regresijske premice, ki se shranijo v spremenljivke $parametri[0]$ in $parametri[1]$. Končna funkcija bo zato imela obliko:

$$proracun = e^{parametri[1]} \cdot e^{parametri[0] \cdot leto}$$

Konkretni vrednosti naših parametrov pa je program izračunal, da sta približno $parametri[0] = 0,079$ in $parametri[1] = -139,55$. Zdaj lahko na naš graf narišemo eksponento funkcijo in si ogledmo kako dobro se prilega podatkom.



Slika 6: Aproksimacija največjega letnega proračuna

S pomočjo aproksimirane funkcije lahko ocenimo največji proračun za film v letu 2018:

$$proracun = e^{-139,55} \cdot e^{0,079 \cdot leto} = e^{-139,55} \cdot e^{0,079 \cdot 2018} \approx 596 \cdot 10^6$$

Po naših izračunih lahko torej pričakujemo film, ki bo imel več kot pol milijarde dolarjev proračuna.

3 Zaključek

Ugotovili smo, da je povprečna ocena vseh filmov precej nad sredino $5,0$ in da ni popolnoma vseeno kdaj je film izdan, če se zanimamo za čim večji dobiček. Prav tako smo ugotovili, da med oceno in dobičkom ni očitne povezave oziroma linearne odvisnosti. To pomeni, da dobra ali slaba ocena gledalcev nima velikega vpliva na dobiček. Pogledali smo tudi kolikšen proračun filmska podjetja razpisujejo za svoje filme. Tu smo opazili in aproksimirali eksponentno rast v največjih letnih proračunih.
Mislim, da je naloga dober primer statistične analize na podatkih iz realnega sveta, saj smo uspeli ugotoviti nekaj lastnosti filmov, ki morda sicer niso povsem očitne.

Viri

[1] Aleksandar Jurišić. Verjetnostni račun in statistika, zapiski [online], Ljubljana 2017. https://ucilnica.fri.uni-lj.si/pluginfile.php/14524/mod_page/content/43/vs1710.pdf, dostopano 3.1.2018

[2] Kaggle [online], https://www.kaggle.com, dostopano 3.1.2018

[3] The Movie Database (TMDb) [online], https://www.themoviedb.org, dostopano 3.1.2018

[4] The Movie Database (TMDb) 5000 Movie Dataset [online], https://www.kaggle.com/tmdb/tmdb-movie-metadata, dostopano 3.1.2018