TELIA LIETUVA | DATA SCIENCE INTERNSHIP TASK

KRISTIJONAS SILIUS

2024 07 02

UŽDUOTIS

- Gauti 3 filmų duomenų failai iš MovieLens:
  - movies_metadata.csv: duomenys apie filmus
  - keywords.csv: filmų raktiniai žodžiai
  - ratings.csv: 700 žiūrovų vertinimai suteikti 9000 filmų skalėje nuo 1 iki 5
- Tikslas atlikti analizuoti duomenis, pateikti išvadų, grafikų, įžvalgų apie kas gali lemti filmų populiarumą.

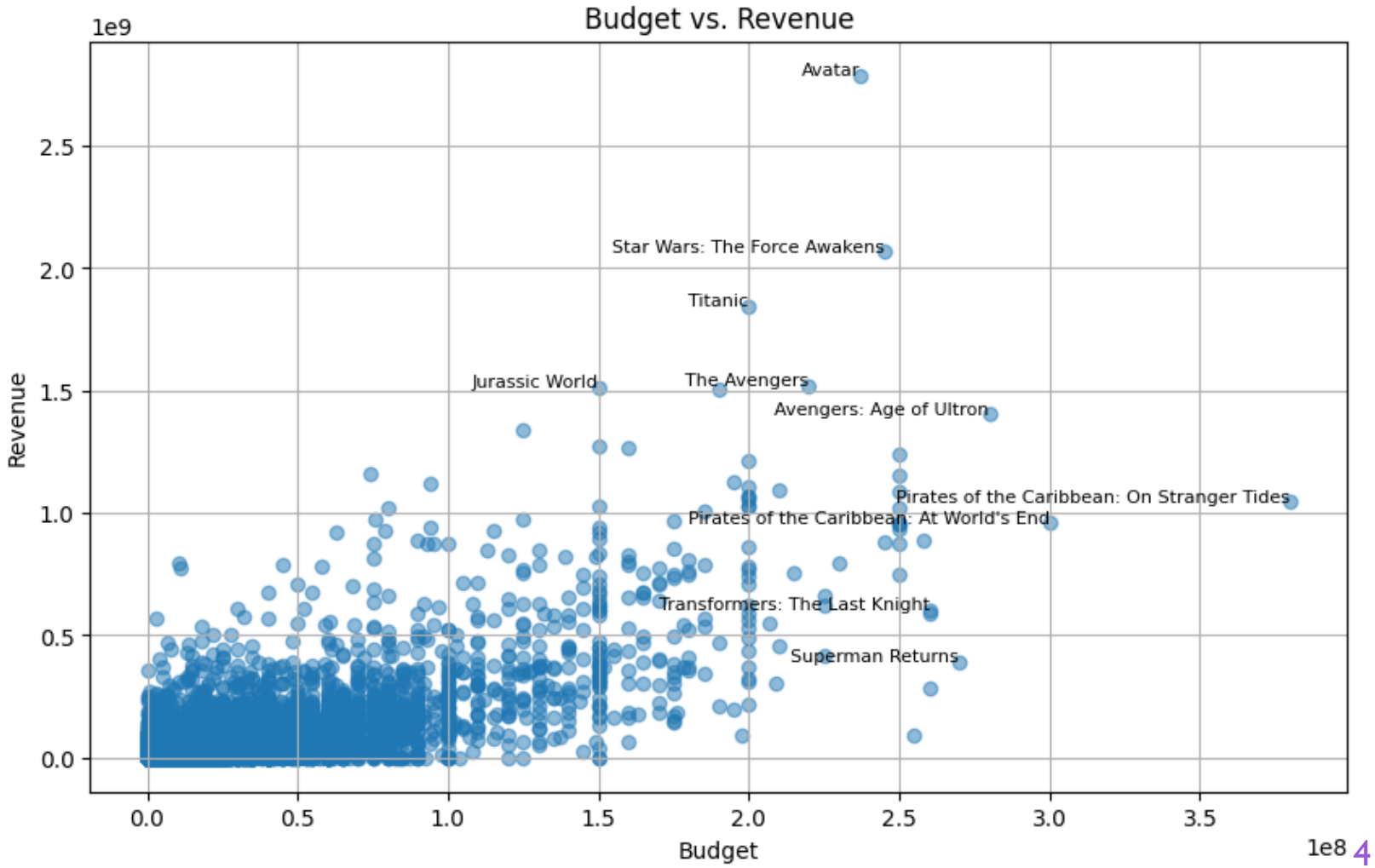
- Rezultatai pateikti šioje prezentacijoje, bet visą Jupyter Notebook kodą galima rasti prisegtą atskirai

REZULTATAI

- Pirmas žingsnis buvo pasitikrinti duomenis
- Ir atlikti data cleaning procesą (visi žingsniai notebook’e)

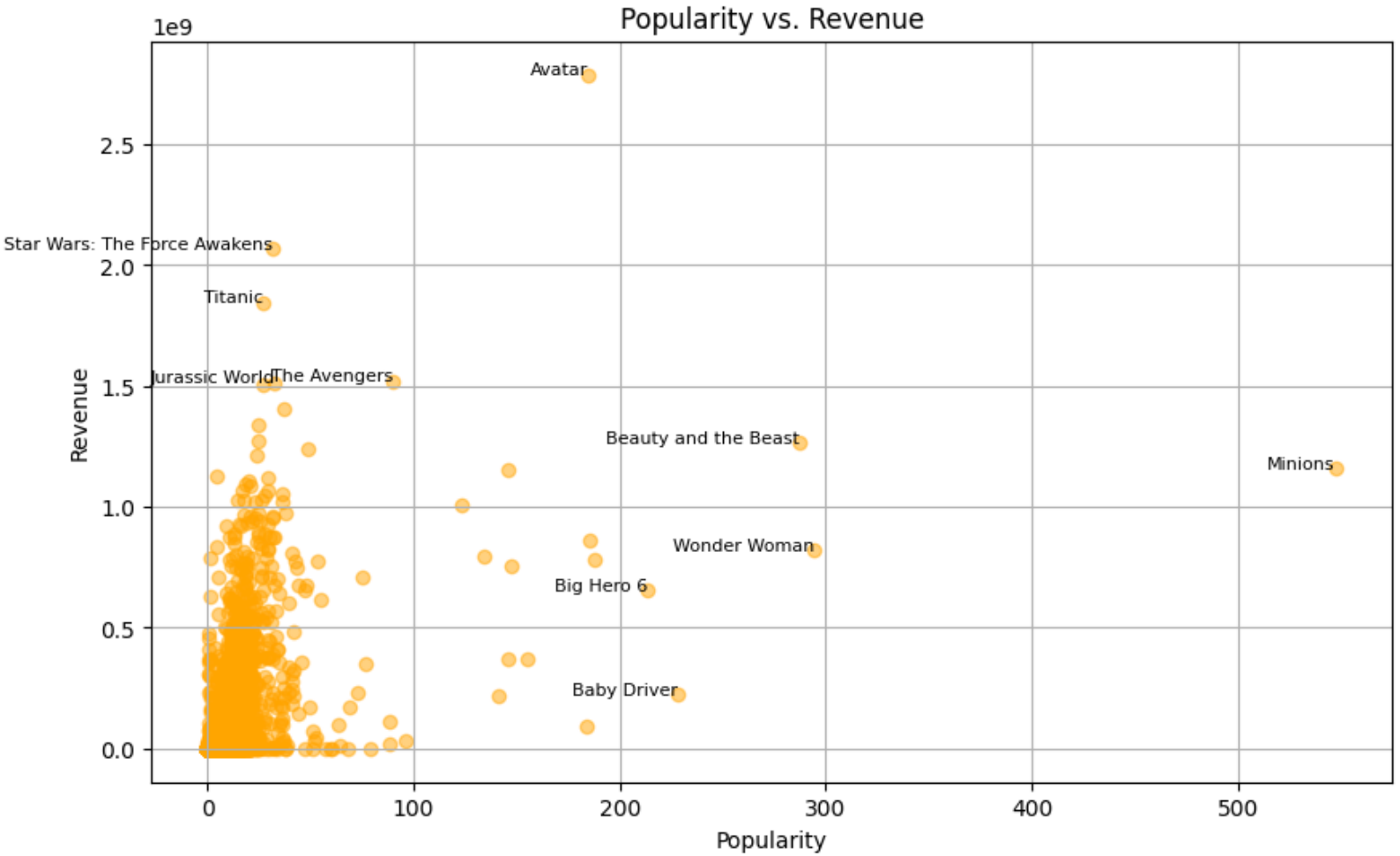
REZULTATAI

- Paruošus duomenis darbui. Pirmas dalykas ką patikrinau buvo kaip filmų pajamos (revenue) priklauso nuo jų biudžeto.
- Matoma tendencija, kad didesnio biudžeto filmai dažnai atneša daugiau pajamų.
- Tačiau yra išskirtinių atvejų kaip Avatar ir Pirates of the Caribbean: On Stranger Tides: kur atitinkamai su mažesniu biudžetu pajamos yra didžiausios ir vice versa.



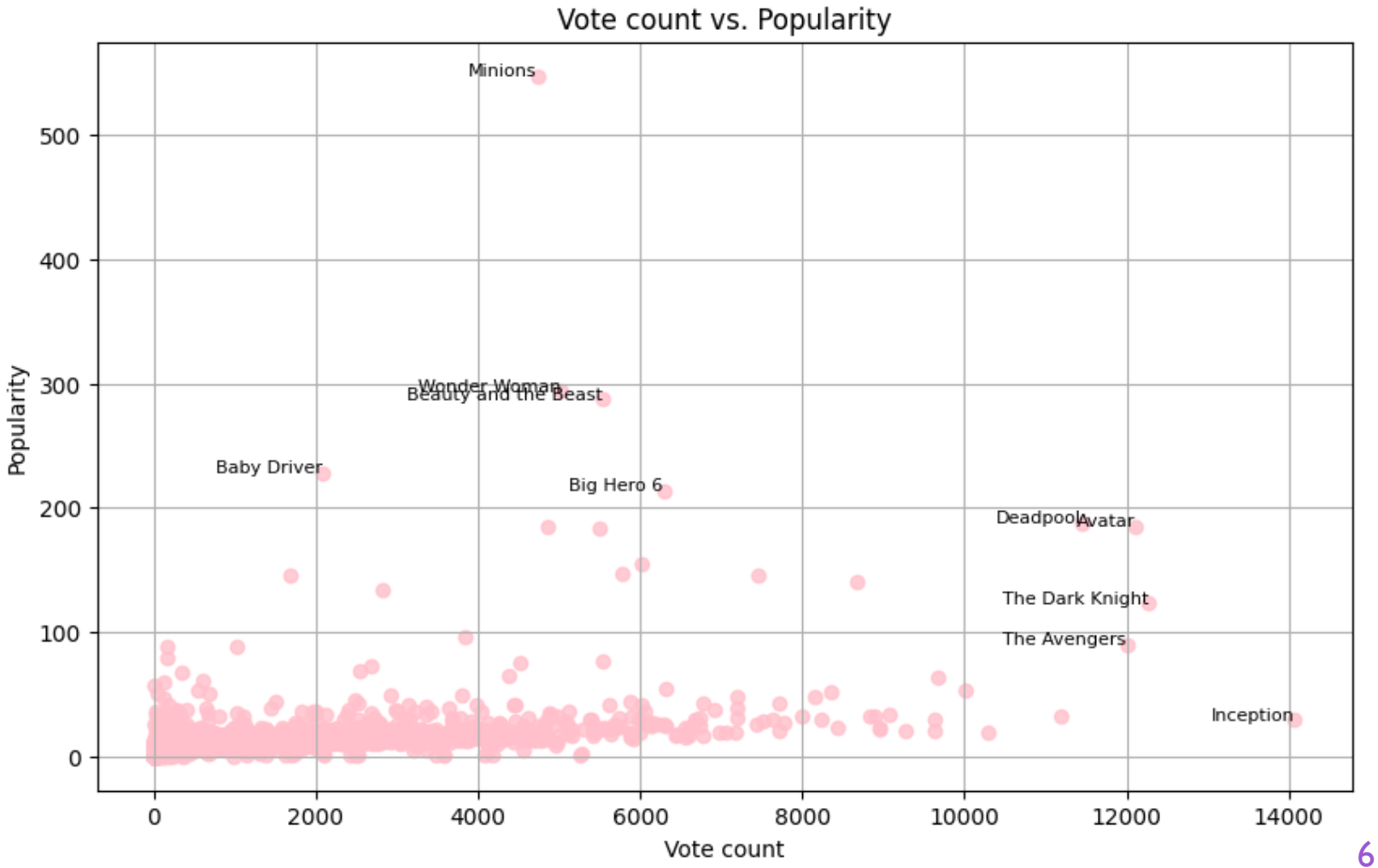
REZULTATAI

- Toliau palyginau pajamas nuo populiarumo.
- Čia tendencija daug „statesnė“, kas indikuoja mažesnę tiesinę koreleciją.
- Vėl išsiskiria Avatar dėl didžiausio pajamų skaičiaus, bet ne tokio didelio populiarumo, o Minions populiarumo reitingas labai aukštas, bet pajamos vidutinės.



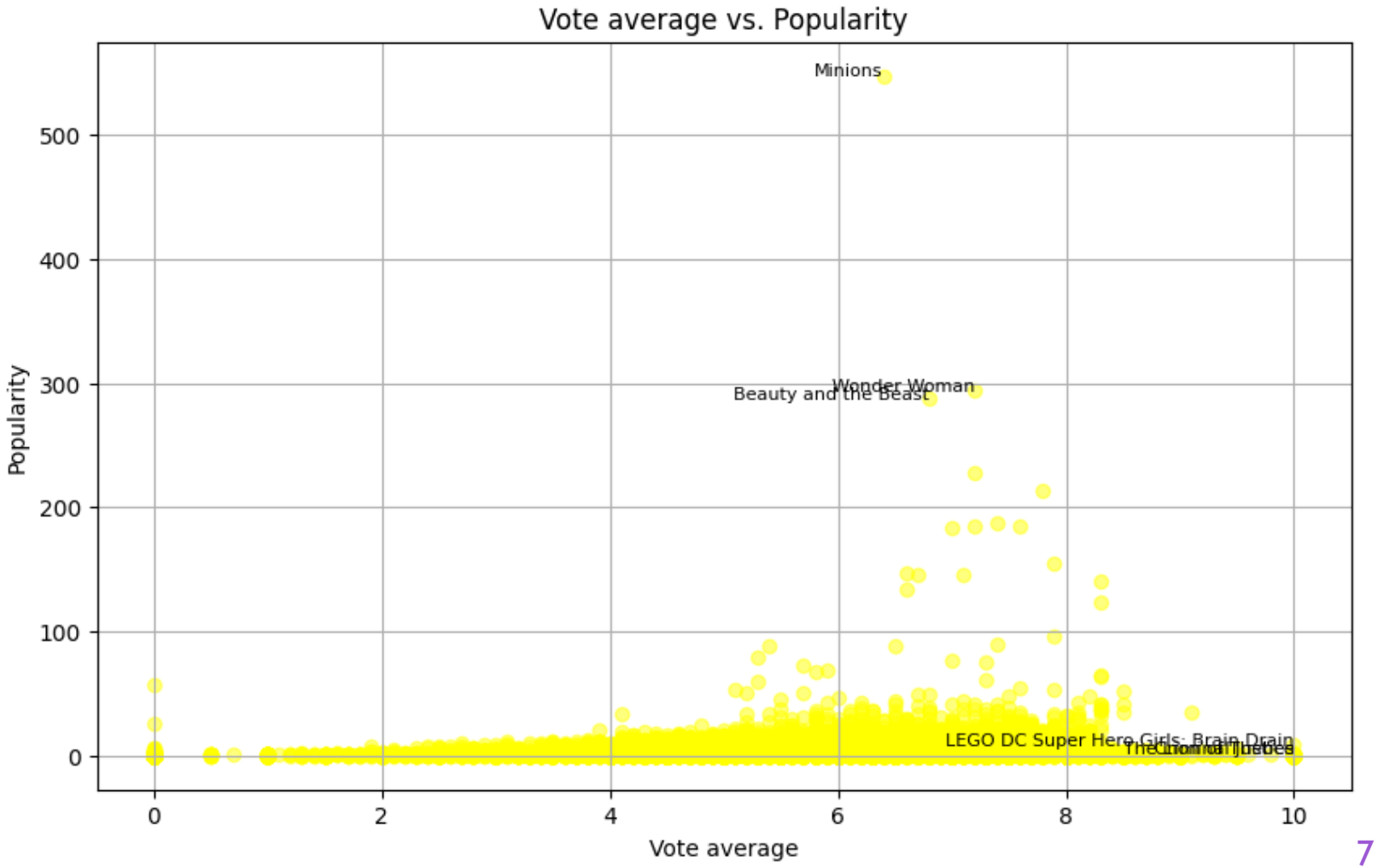
REZULTATAI

- Palygintas filmų populiarumas nuo reitingų skaičiaus.
- Tendencingumo didelio nėra.
- Išsiskiria Minions – didelis populiarumas vidutiniškas reitingų skaičius (filmas vaikams – tikriausiai mažiau linkę reitinguoti). Inception ne didelis populiarumas, bet labai daug reitingų (gal galima sakyti, kad thought provoking filmas, apie kurį žmonės nori išreikšti nuomonę).



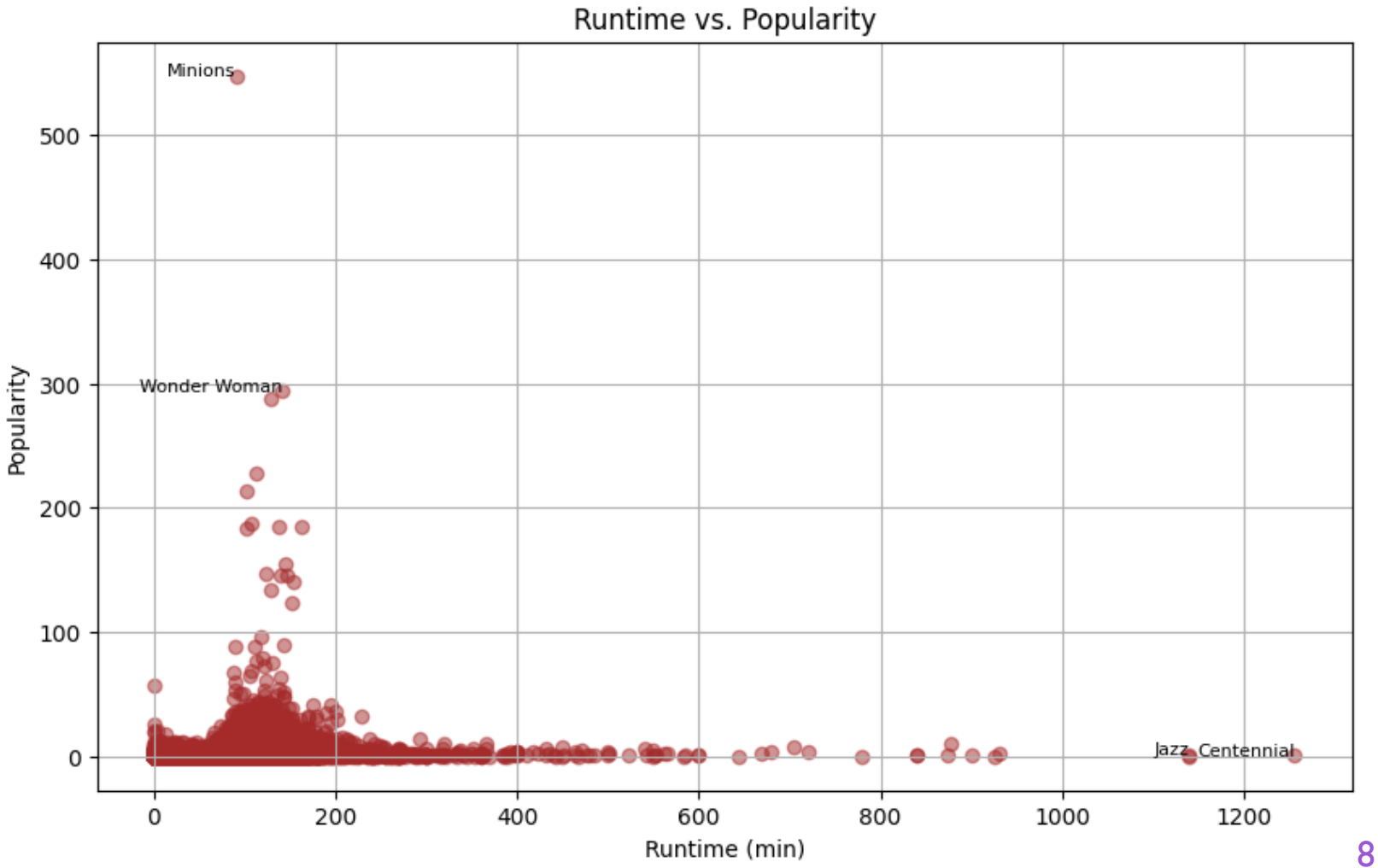
REZULTATAI

- Palygintas filmų populiarumas nuo jų reitingų.
- Tendencingumas rodo, kad filmai įvertinti ~6.5-8.5 intervale yra populiariausi.
- Išsiskiria ir šioks toks populiarumo šuolis prie filmų įvertintų 0.



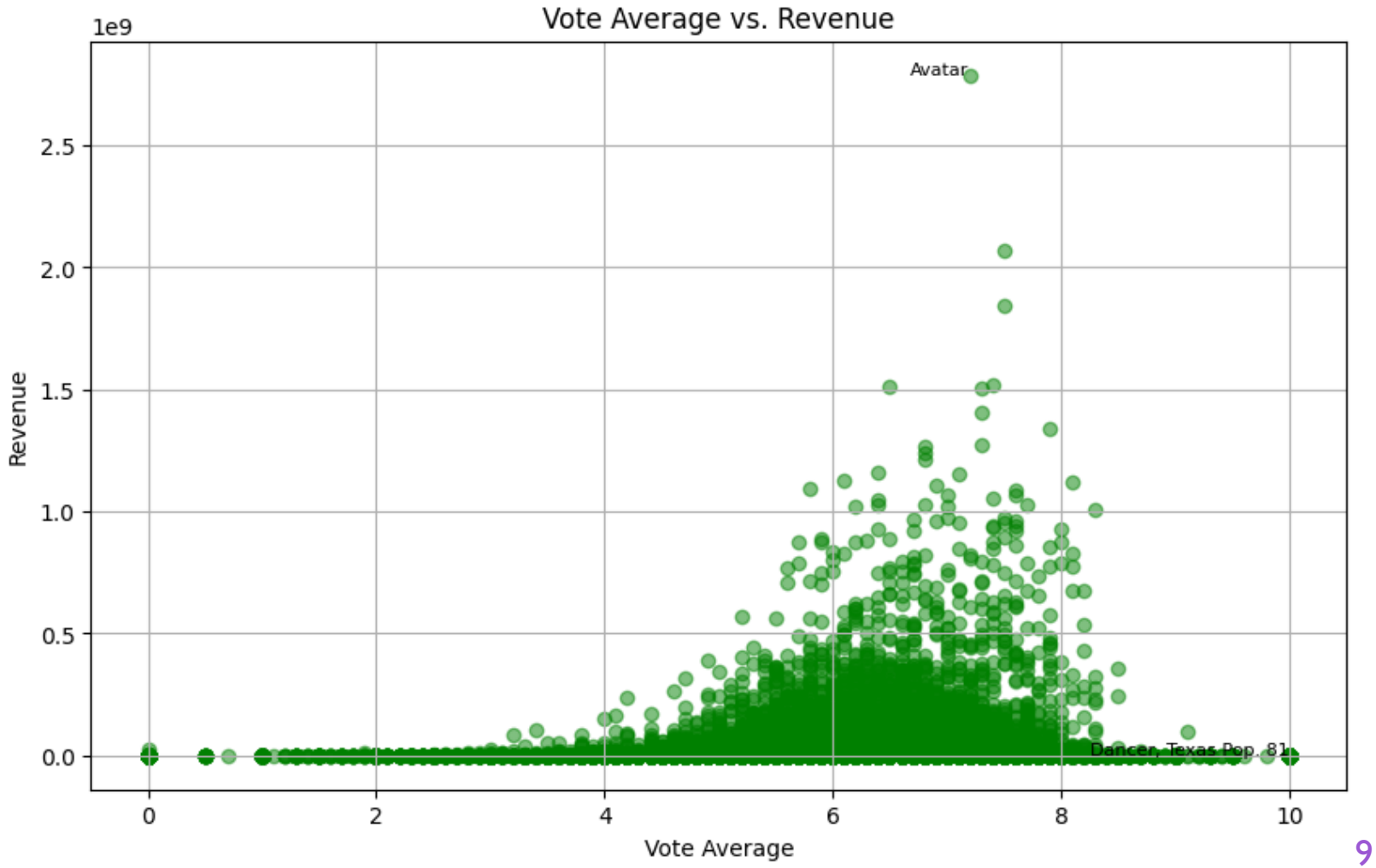
REZULTATAI

- Palygintas filmų populiarumas nuo trukmės (runtime in minutes).
- Tendencingumas rodo, kad filmai, kurie trunka intervale nuo ~100 min. iki 200 min. yra populiariausi.
- Centennial trukmė labai ilga, bet patikrinus supratau, kad jie traktuoja, kai kuriuos TV shows su daug epizodų kaip filmus, tai paaiškina trukmę.



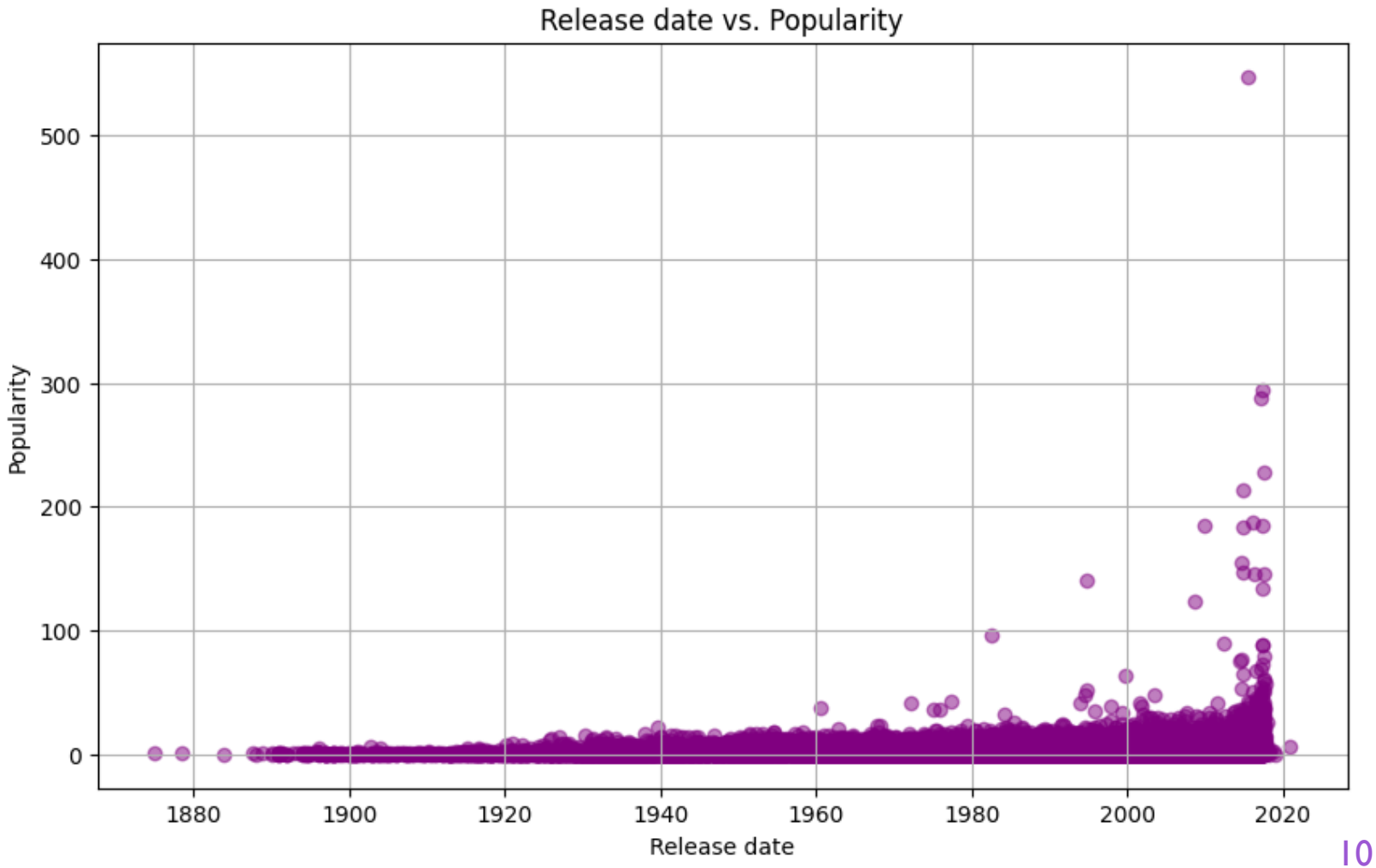
REZULTATAI

- Palygintos filmų pajamos nuo jų reitingo.
- Tendencingumas rodo, kad pelningiausi filmai įvertinti intervale nuo ~5 – 8.5 (su didžiausiu piku prie reitingų tarp 7 – 8).
- Beveik normalus, gausinis skirstinys matomas nurodytuose intervaluose.
- Įdomu, kad geriausiai įvertinti filmai nėra pelningi (kaip ir logiška turėti vidutinį 10 įvertinima indikuoja labai mažą balsų skaičių).
- Uždėjau mažiau labels, nes labai persikloja ir nesimato.



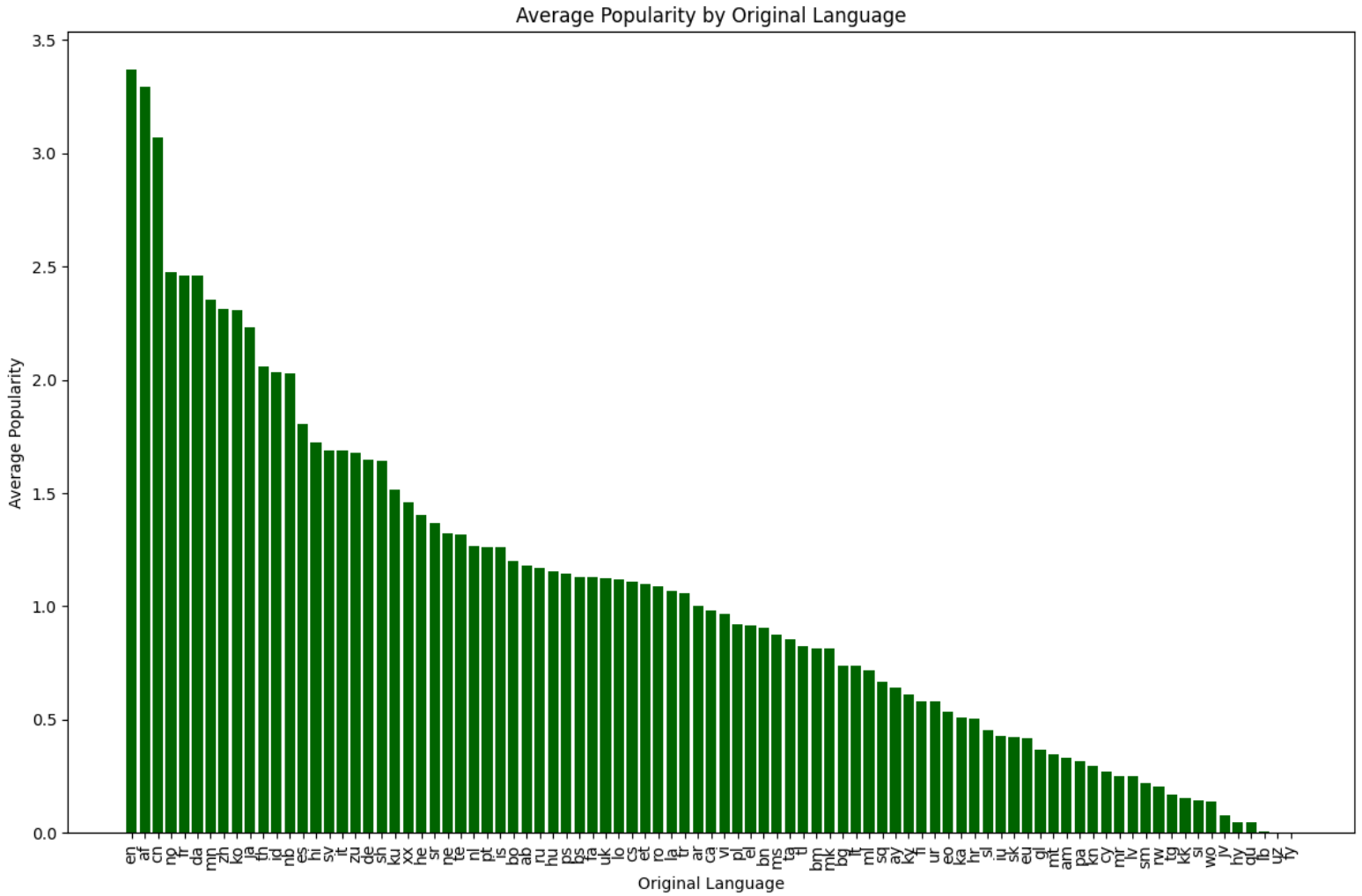
REZULTATAI

- Palygintas filmų populiarumas nuo jų išleidimo datos.
- Tendencija, jog naujesni filmai daug populiaresni.
- Labai logiška – dabar daugiau filmų, kuriamų, lengviau prieinamas jų žiūrėjimas, geresni efektai, technologijos ir pan.



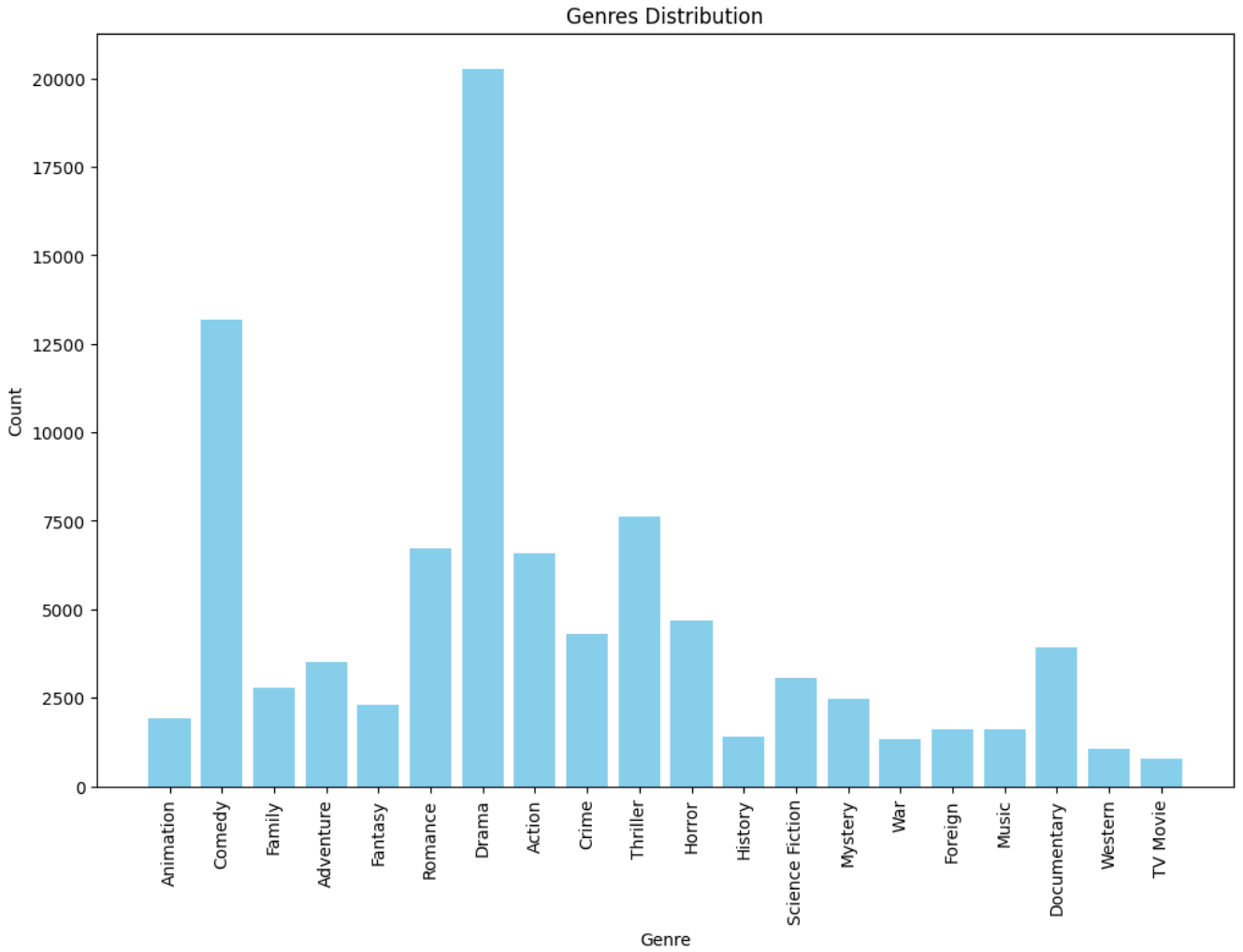
REZULTATAI

- Palygintas filmų populiarumas nuo jų orginalo kalbos.
- Dominuoja anglų, afrikiečių ir kinų kalbomis kurti filmai.
- (cn – Cantonese ir dažnai panašu atitinka Hongkongą, o zh – Mandarin dažnai atitinka Kiniją)
- (af – panašu atitinka Pietų Afriką)



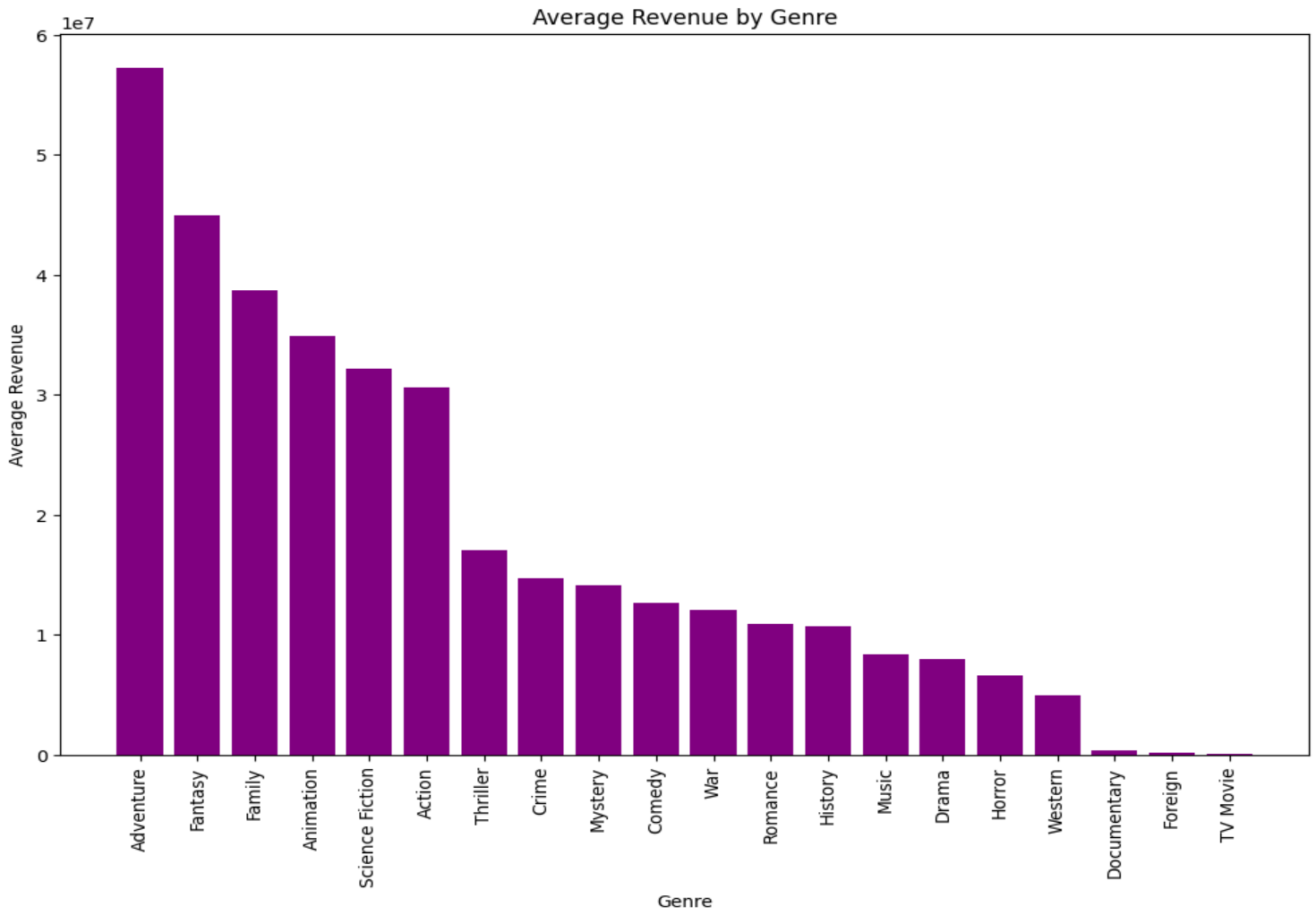
REZULTATAI

- Atvaizduotas filmų žanrų pasiskirstymas.
- Daugiausia yra dramos, komedijos ir trilerių.



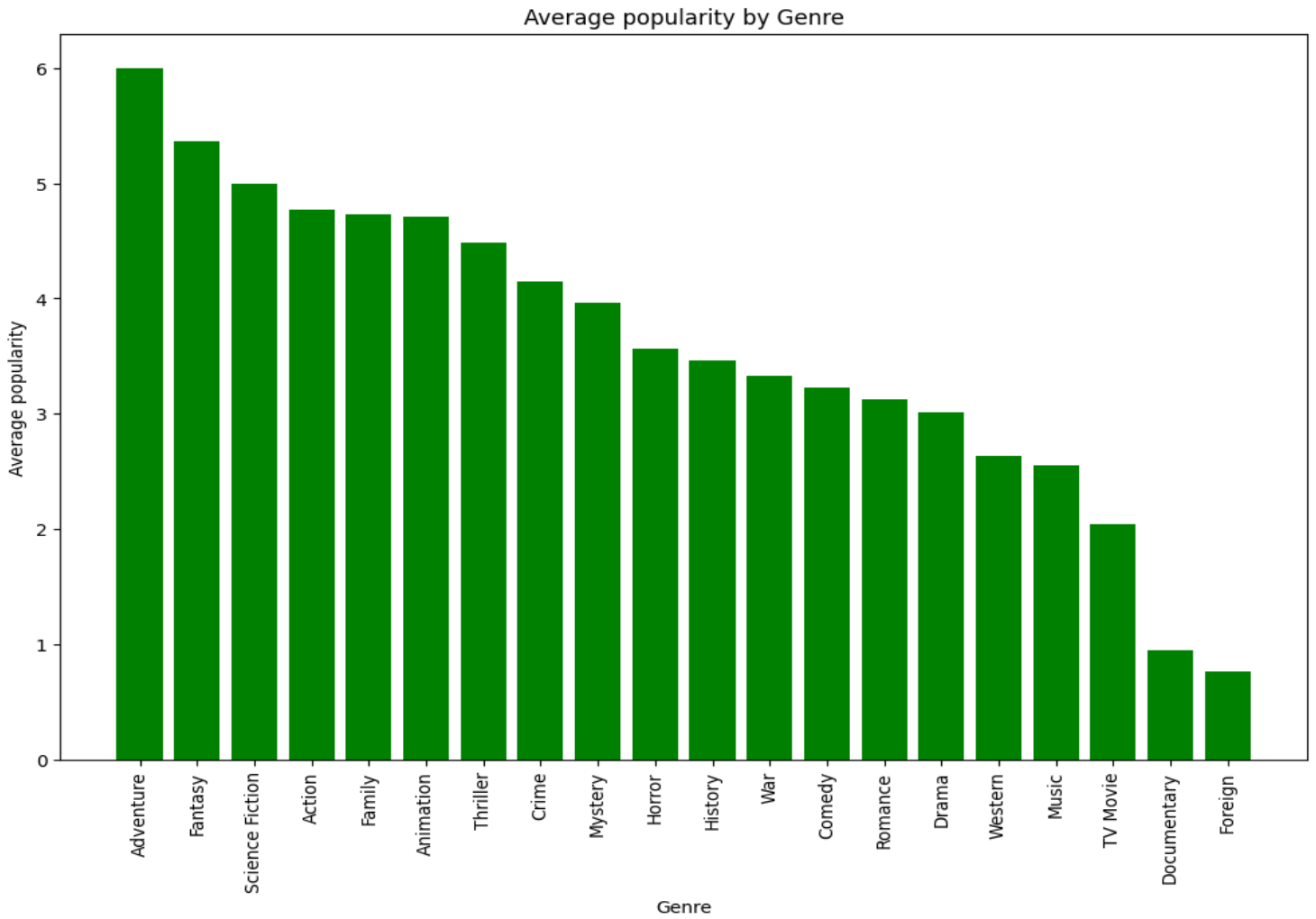
REZULTATAI

- Palygintos vidutinės filmų pajamos kiekvienam žanrui.
- Daugiausia pajamų sukuria nuotykių, fantastiniai ir šeimyniniai filmai (dažnai gali perskikloti adventure ir fantasy arba family ir animation ir pan.)
- Tuo tarpu western, dokumentiniai, užsienio uždirba mažiausia. (nesu tikras kodėl TV Movie yra žanras, bet irgi mažai pajamų atnešantis)



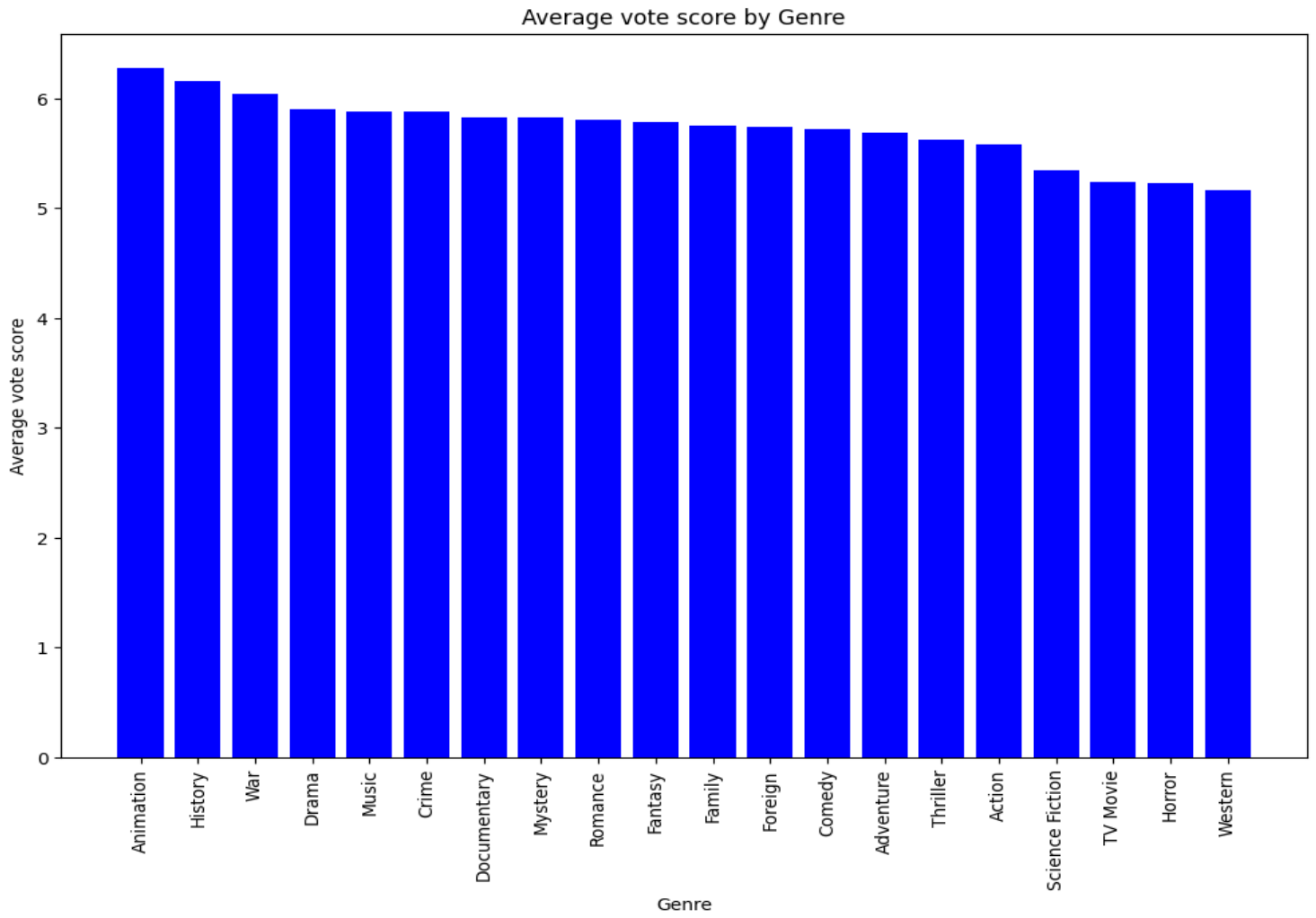
REZULTATAI

- Palygintas vidutinis filmų populiarumas kiekvienam žanrui.
- Populiauriausi beveik visi tie patys žanrai, kaip ir daugiausiai uždirbantys (nuotykių, fantastiniai, mokslinės fantastikos, veiksmo, šeimyniniai).
- Tuo tarpu dokumentiniai, užsienio ir TV movie uždirba nedaug ir nėra populiarūs.



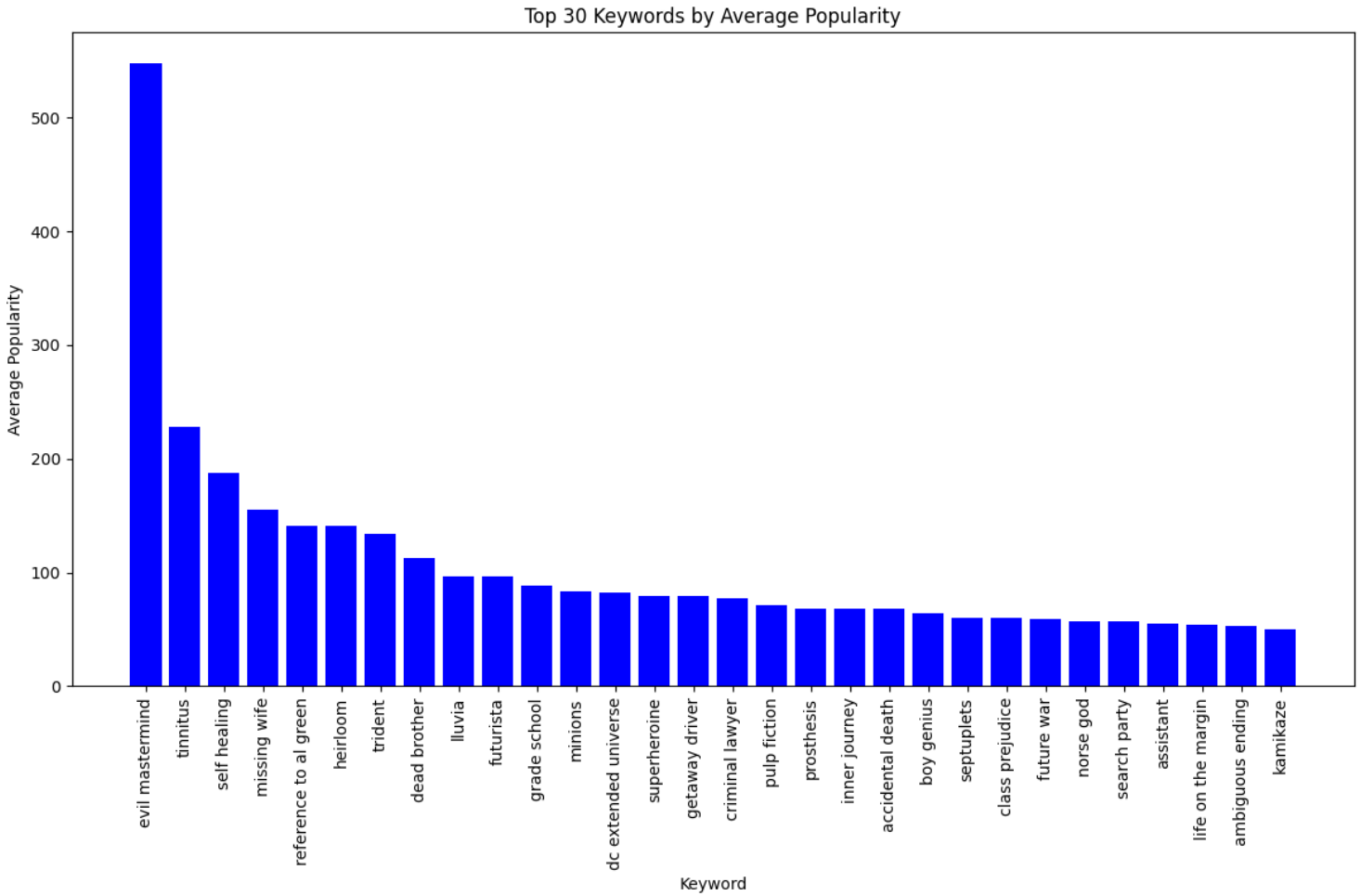
REZULTATAI

- Palygintas vidutinis filmų reitingas kiekvienam žanrui.
- Čia geriausiai įvertinti animaciniai, istoriniai ir karo filmai.
- Bet visi reitingai žanrams turi uniform distribution formą ir visi reitingai ganėtinai panašūs.



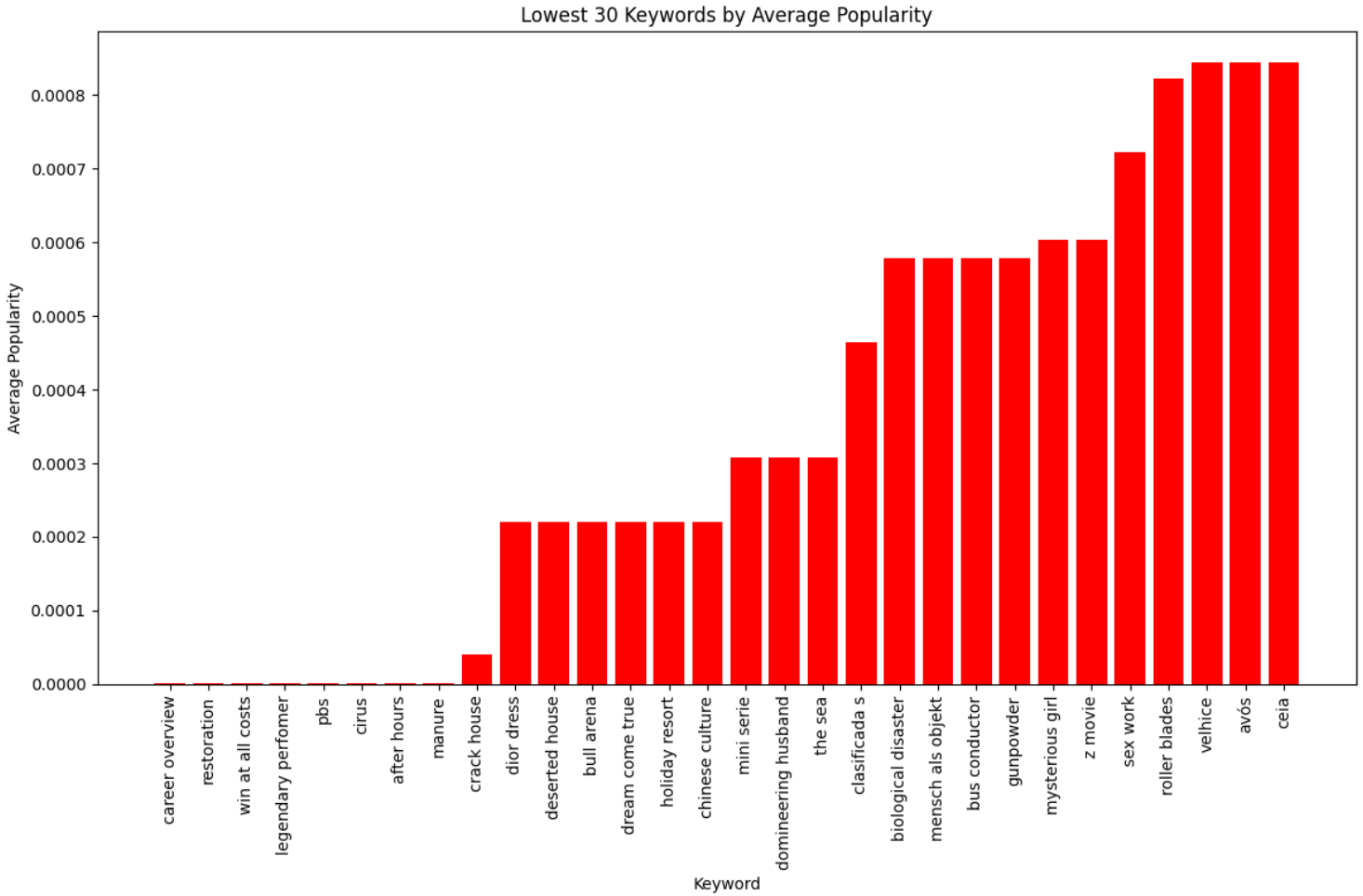
REZULTATAI

- Apjungiau keywords ir movies_metadata failus.
- Lyginama filmų populiarumas jų raktožodžiams. Pateikti top 30 raktožodžių, kuriuos turi populiariausi filmai.
- Didelis populiarumo šuolis „evil mastermind“ raktožodį turintiems filmams.



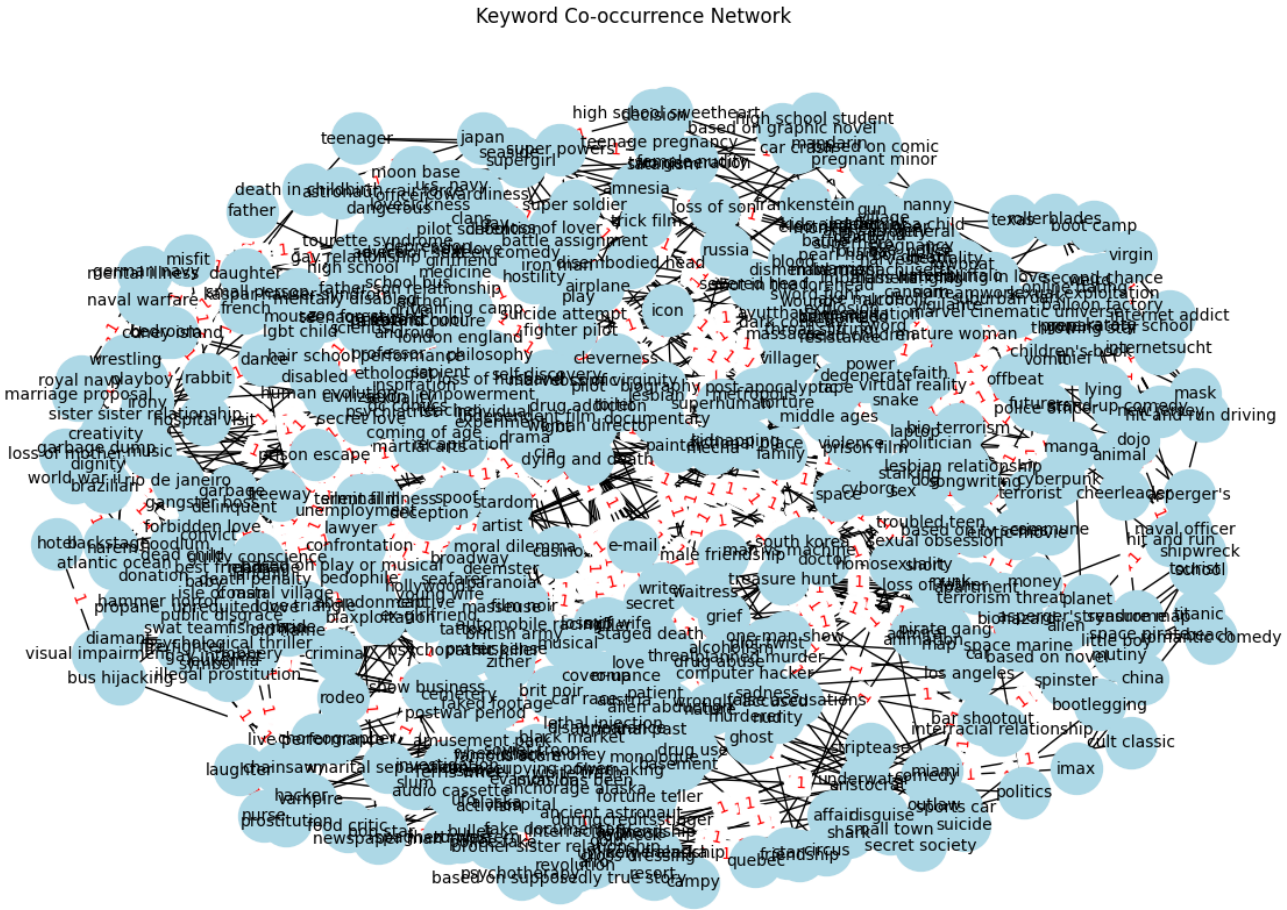
REZULTATAI

- Ekvivalenčiai pateikiu top 30 raktažodžių mažiausio populiarumo filmams.
- Atsiranda daug ne angliškų žodžių ir labai specifiškų raktažodžių kaip „bull arena“.



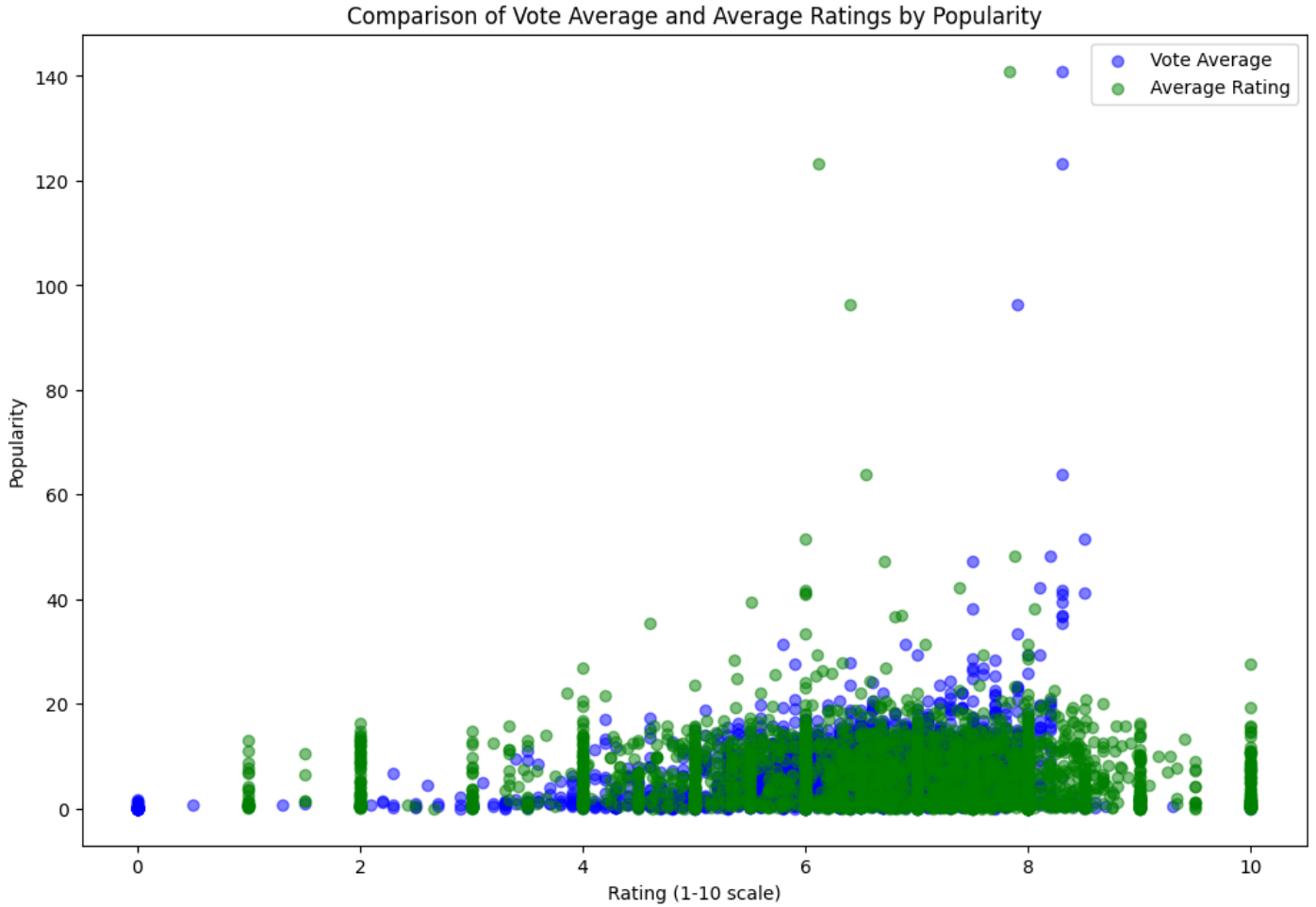
REZULTATAI

- Mėginau apjungti filmų co-occurence tinklą, bet deja, net su mažu filmų sample size (200) vaizdas labai negražus ir neperskaitomas.



REZULTATAI

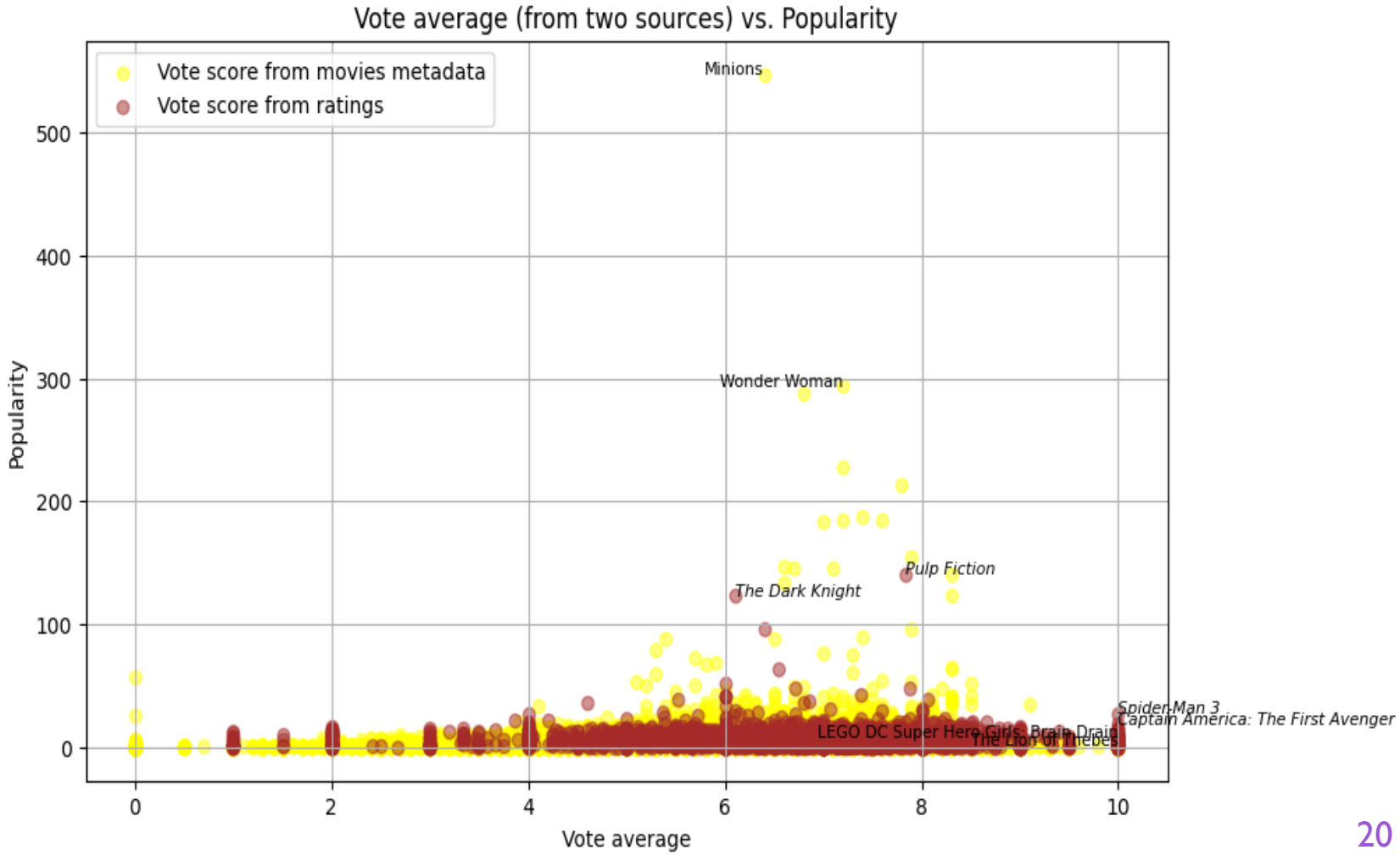
- Apjungiau ratings ir movies_metadata failus (normalizuuoti ratings x2, kad skalė būtų iš 10 abiejuose datasets).
- Žali taškai indikuoja reitingus iš ratings.csv failo, mėlyni iš movies_metadata.
- Tendencingumai ganėtinai panašūs.
- Daug daugiau žalių 10/10 retingų, spėčiau, kad žmonės labiau linkę duoti 5/5 nei 10/10. Šiuo atveju 5/5 pavirto į 10 normalizavus.



REZULTATAI

- Raudoni taškai indikuoja reitingus iš ratings.csv failo, geltoni iš movies_metadata.
- Ratings labels pasukti į dešinę ir *italic*.

- Daug raudonų 10/10 retingų, spėčiau, kad žmonės labiau linkę duoti 5/5 nei 10/10. Šiuo atveju 5/5 pavirto į 10 normalizavus.



REZULTATAI

- Stengiausi sukurti paprastą rekomendacinę sistemą pagal filmų raktažodžius ir vartotojo filmų įvertinimus (detali info basic_recommendation_system.ipynb)
- Pritaikau cosine similarity paimtiems vartojo filmų raktažodžiams (scaled pagal jų filmo reitingą) ir pagal tai surandama rekomenduojamų filmų.
- Labai netobula sistema, daug ką galima taisyti, bet šioks toks bandymas.
- Šiuo atveju duodama Toy Story (862): su reitingu 5, Grumpier Old Men (15602) rating 4, Heat (949) rating 3.
- Ir sistema rekomenduoja: Grumpier Old Men, Star Trek IV, Bad Santa ir My sister's Keeper.

```
# USE THE SYSTEM

# user_ratings = {movieId: rating, ...}
user_ratings = {862: 5, 15602: 4, 949: 3, 168: 4} # example values to test
user_profile = create_user_profile(user_ratings)
recs = generate_recommendations(user_profile)
print(recs)
```

```
[15602   168 10024 15578  5669 10147 17393  7993]
```

```
# recommended movie IDs link to their index values
Codiumate: Options | Test this function
rec_indices = movies_df.index[movies_df['id'].isin(recs)]

# check the titles of the recommended movies based on index with matchi
recommended_movies = movies_df.loc[rec_indices, 'title']

# will print out the movie title if the id exists in the file, the numb
print(recommended_movies)
```

```
2                    Grumpier Old Men
1330     Star Trek IV: The Voyage Home
6816                         Bad Santa
13890               My Sister's Keeper
Name: title, dtype: object
```

IŠVADOS

- Pavyko ganėtinai sėkmingai paanalizuoti filmų duomenis.
- Tendencijos rodo, kad populiariausi filmai dažnai bus įvertinti tarp 6.5 – 8.5.
- Populiarumą lemia:
  - Filmų trukmė, optimali nuo ~100 min. iki 200 min.
  - Išleidimo data – naujesni filmai daug populiaresni už senus.
  - Orginalo kalba – anglų, afrikiečių, kinų, norvegų išsiskiria.
  - Žanrai – nuotykių, fantastiniai, mokslinės fantastikos, veiksmo, šeimyniniai, animaciniai.
- Populiarūs filmai turi šiuos raktožodžius: evil genius, tinnitus, self healing, missing wife, reference to al green.