Mid-bootcamp project Movie analysis

DATA ANALYTICS | IRONHACK

Ralph Ward

Context

Underlying factors of what makes a movie popular

—

Movies recommenders -

The age of personalization

Source: AI Midjourney prompt

Data Source

- My watched movies:
  Pinterest scraper ~ 1,3k movies
- IMDb database:
  Kaggle API ~ 250k movies
- Oscar & Golden Globes awards:
  Kaggle csv - 1920s to 2020s

Data Limits

- Empty values for votes & rankings on ~30%
- Movie scorer, weighted average & time decay factor
- Huge dataset: encoding difficulties

EDA Questions

Most popular genres?

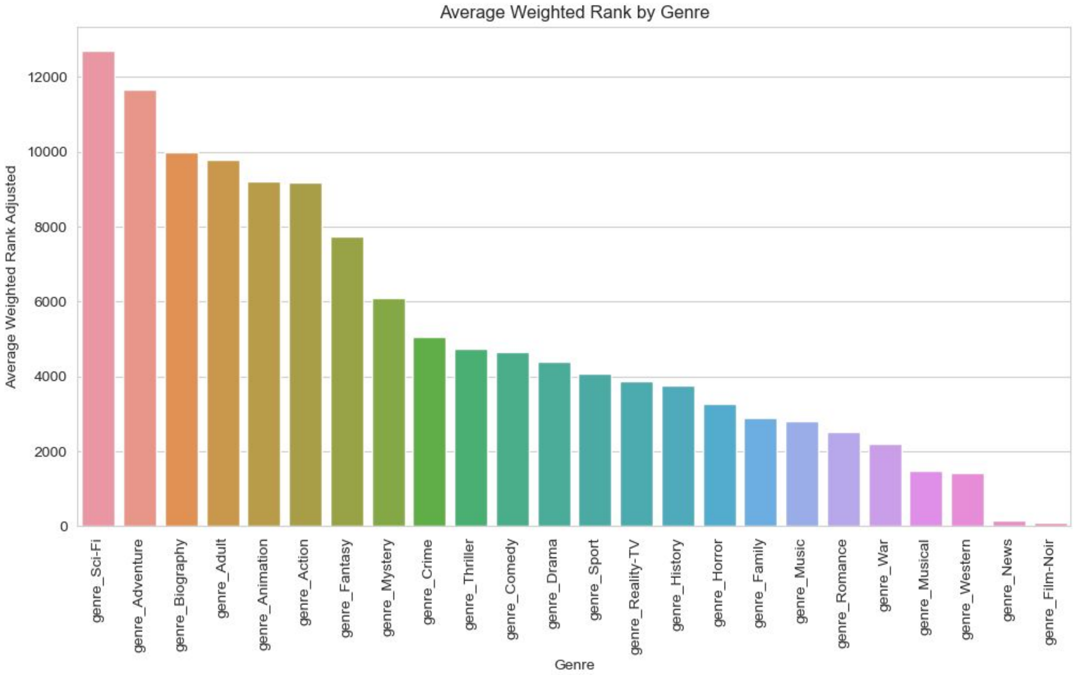
The best actors?

Hypothesis:

Do best actors influence the movie ratings?

Best directors?

Best movies?



🧐 Surprisingly for me “Biography” is high and “Comedy” is low

🧐 “Westerns” are definitely dying off

⚠️ “Drama” is often bundled with other genres

EDA Questions

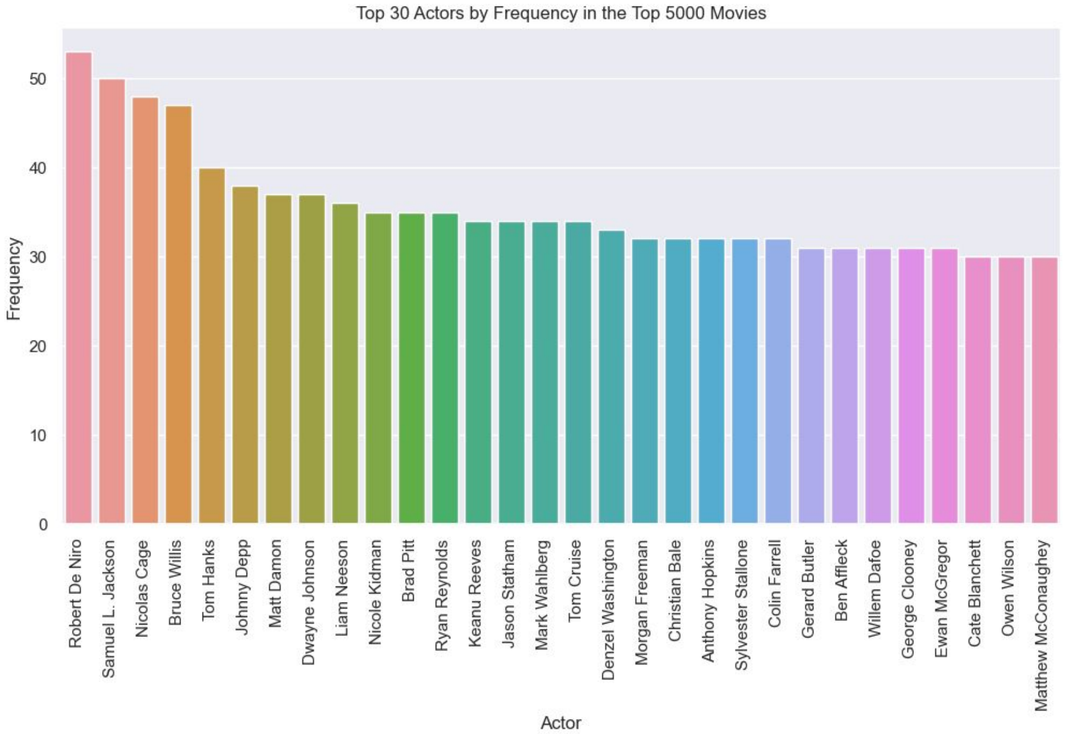
Most popular genres?

The best actors?

Hypothesis:
Do best actors influence the movie ratings?

Best directors?

Best movies?



🧐 As expected, but no Al Pacino?

⚠️ Only 2 actresses!

IRONHACK BOOTCAMP

EDA Questions

Most popular genres?

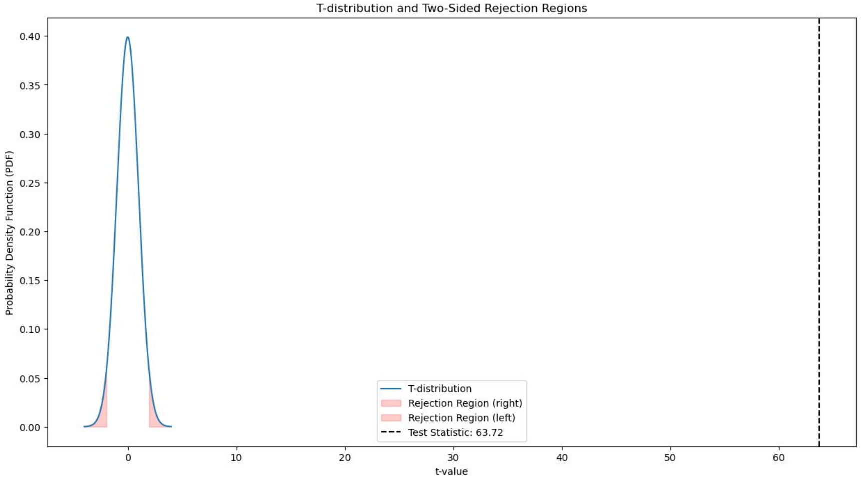
The best actors?

Hypothesis:
Do best actors influence the movie ratings?

Best directors?

Best movies?

H0: No influence on movie ratings
H1: Significant influence



```
Confidence
level: 0.95

T-stat:
63.72188

P-value:
0.0
```

🧐 Reject H0, strong significance that there is an influence

⚠️ Actors needed to have award *before* movie casting

EDA Questions

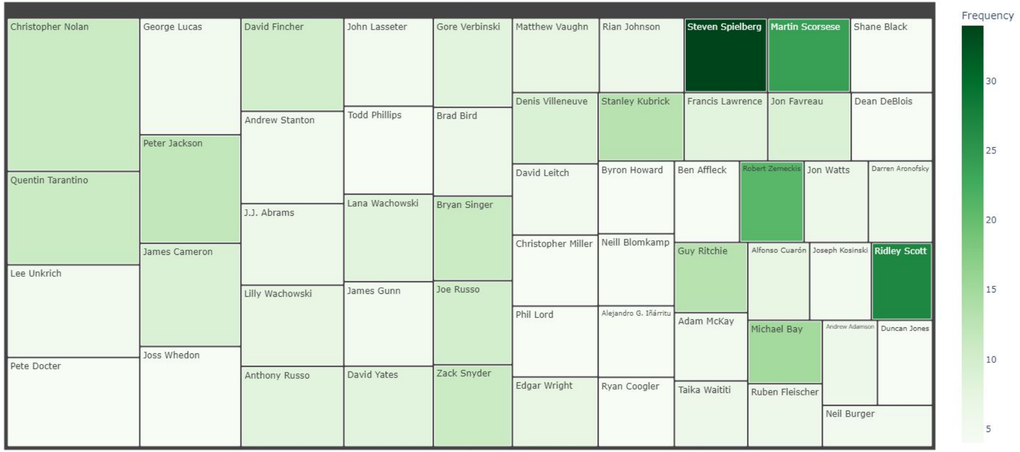
Most popular genres?

The best actors?
Hypothesis:
Do best actors influence
the movie ratings?

Best directors?

Best movies?

Top Directors by average top rankings



🧐 Surprising that Scorsese is low and Ben Affleck is there
🧐 Some directors score high with not many movies released
⚠️ Some directors work together (Russo brothers)

EDA Questions

Most popular genres?

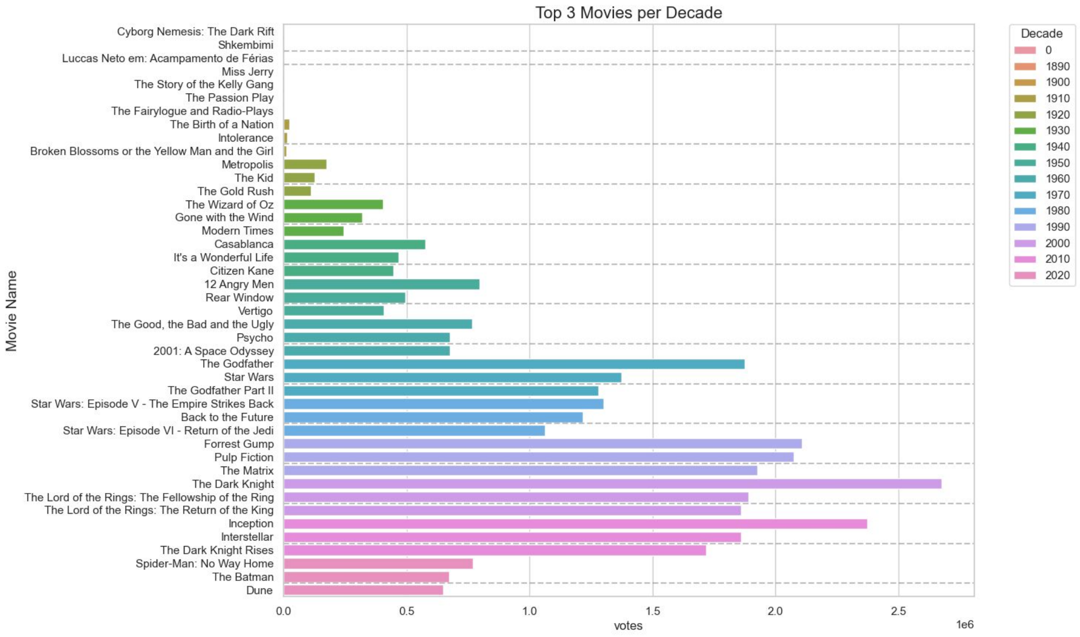
The best actors?

Hypothesis:

Do best actors influence the movie ratings?

Best directors?

Best movies?



🧐 No Titanic or Avatar?

⚠️ More votes for movies released after the Internet!

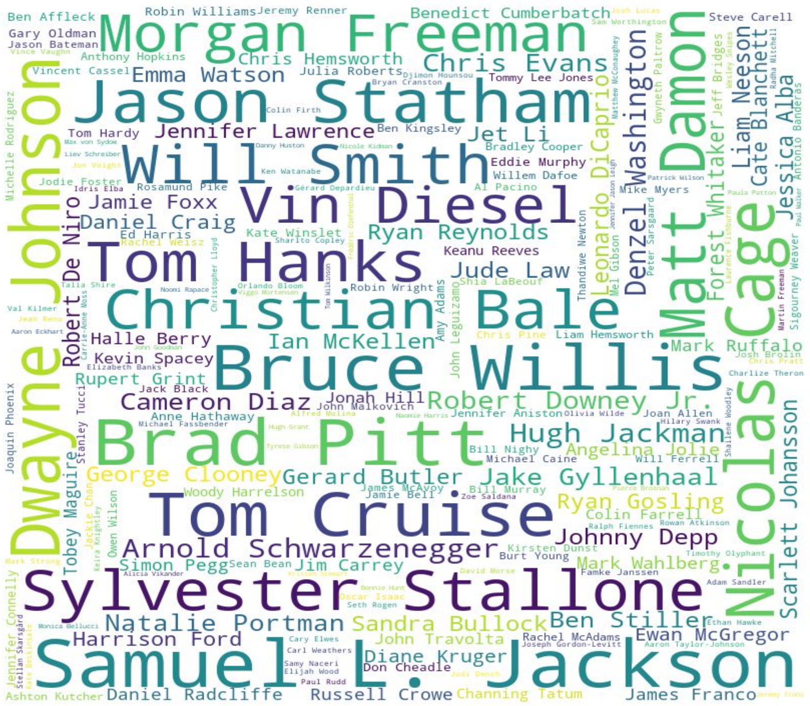
IRONHACK BOOTCAMP

EDA Questions

My preferred actor?

Average rating?

Taste over time?



🧐 I’m boring, this is testosterone fueled!

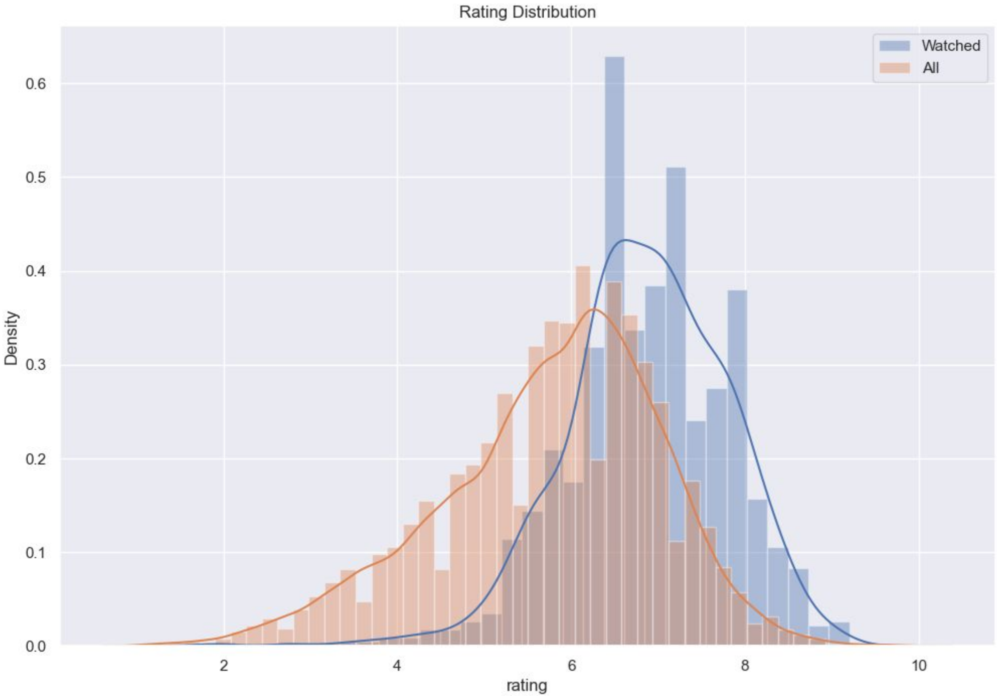
⚠️ Takes into account second & third actors too

EDA Questions

My preferred actor?

Average rating?

Taste over time?



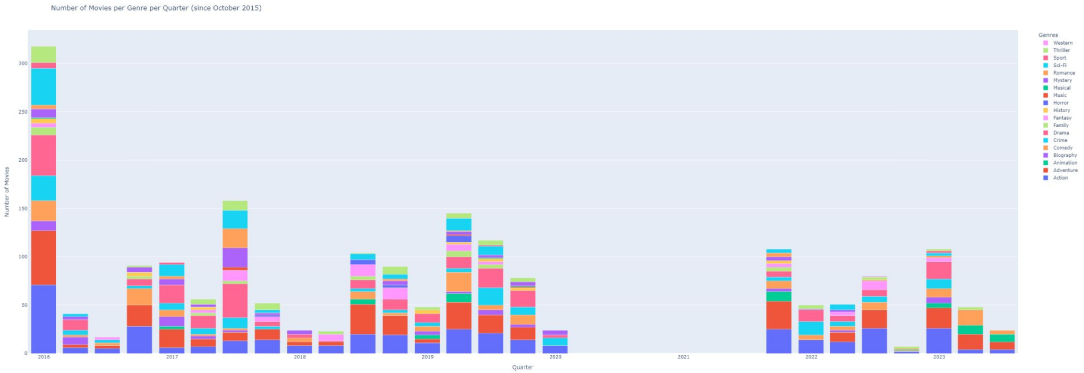
🧐 I have better than average taste!

EDA Questions

My preferred actor?

Average rating?

Taste over time?



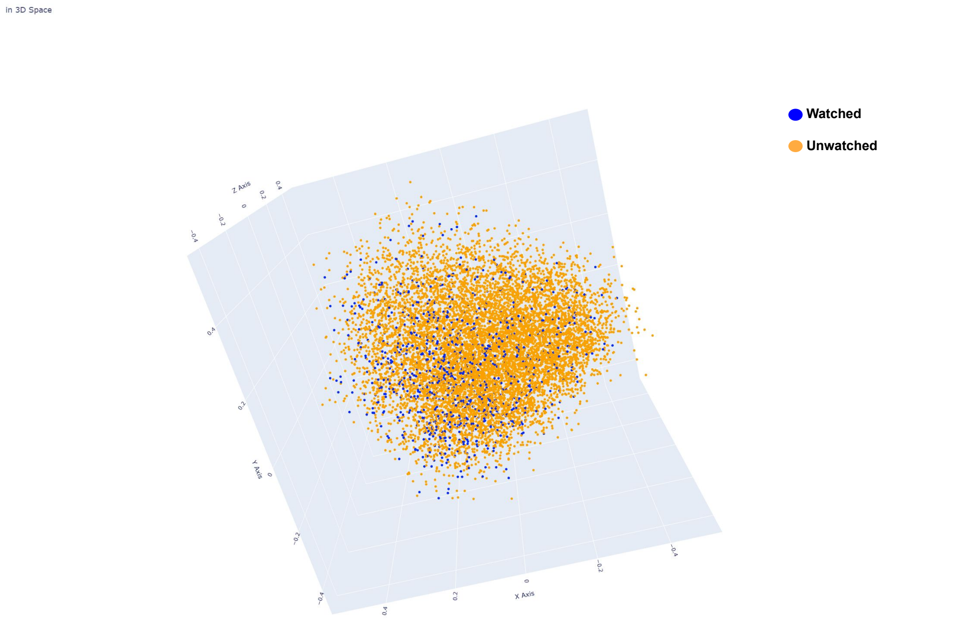
🧐 No good insights

⚠️ I was not consistent with adding movies

⚠️ Too many genres are bundled

Bonus: Movie recommender

Content based system using the movie plot description



- ⚠️ Text to numeric is challenging
- ⚠️ PCA limits with features
- ⚠️ No clear clusters
- ⚠️ Movie similarity output is poor

Concepts: Tokenization, Vectorization, Embeddings, Stopwords, Cosine Similarity, PCA

Conclusion

- Some surprising preferences in genres
- Scoring system identifies best actors & directors
- Significant influence of top actors on movie ratings

Moving Forward

- Additional data sources? Gross revenue etc.
- Refining methodology the the recommender (genre and actors)

Today, personalization is vital and exploring factors shaping individual preferences to tailor the content accordingly is key

IRON
HACK

THANKS !

GitHub LinkedIn