Predicting Global Box Office Revenue

Springboard Capstone #2 Project Report

Maddi Ross

What makes a blockbuster?

Using preproduction movie features to predict box office revenue

Use cases:

- Predicted returns of a budget increase
- Value of producing on-site in a given country, adding additional languages
- Evaluating proposed films for profitability
- Establishing the value of investments in particular marketing features

What makes a blockbuster?

Using preproduction movie features to predict box office revenue

Use cases:

- **Predicted returns of a budget increase**
- Value of producing on-site in a given country, adding additional languages
- Evaluating proposed films for profitability
- Establishing the value of investments in particular marketing features

What makes a blockbuster?

Using preproduction movie features to predict box office revenue

Use cases:

- Predicted returns of a budget increase
- **Value of producing on-site in a given country, adding additional languages**
- Evaluating proposed films for profitability
- Establishing the value of investments in particular marketing features

What makes a blockbuster?

Using preproduction movie features to predict box office revenue

Use cases:

- Predicted returns of a budget increase
- Value of producing on-site in a given country, adding additional languages
- **Evaluating proposed films for profitability**
- Establishing the value of investments in particular marketing features

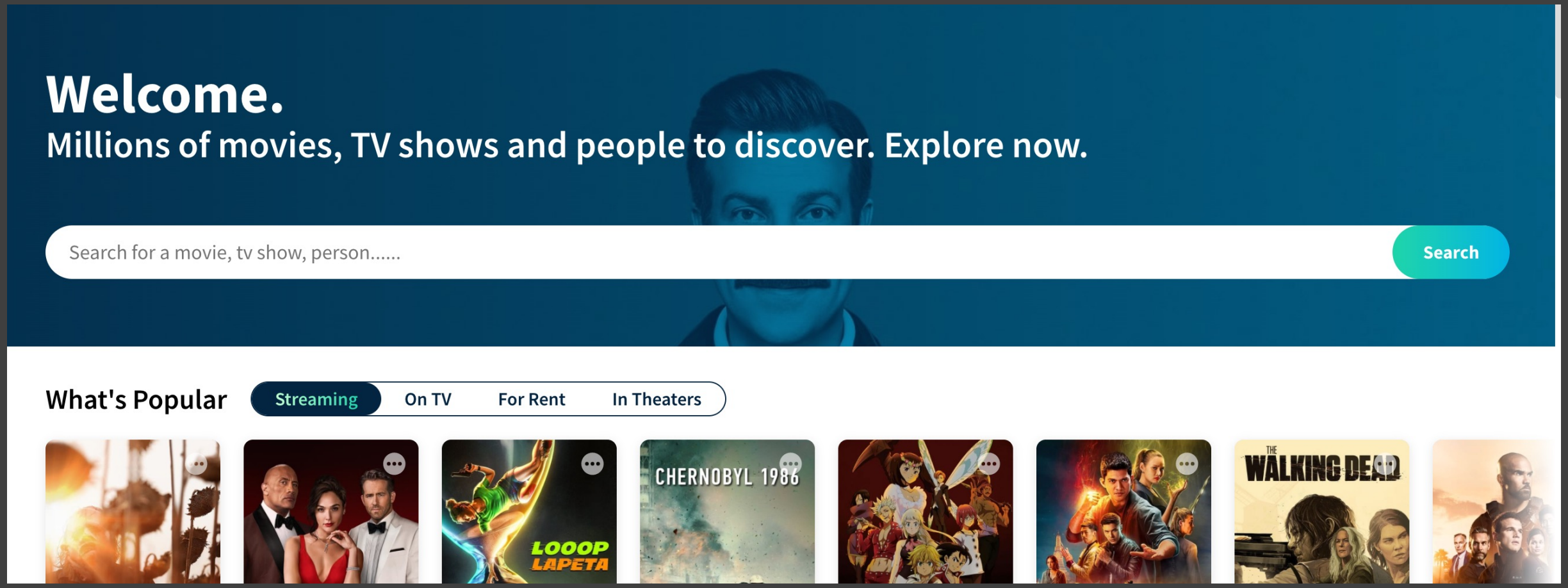
What makes a blockbuster?

Using preproduction movie features to predict box office revenue

Use cases:

- Predicted returns of a budget increase
- Value of producing on-site in a given country, adding additional languages
- Evaluating proposed films for profitability
- **Establishing the value of investments in particular marketing features**

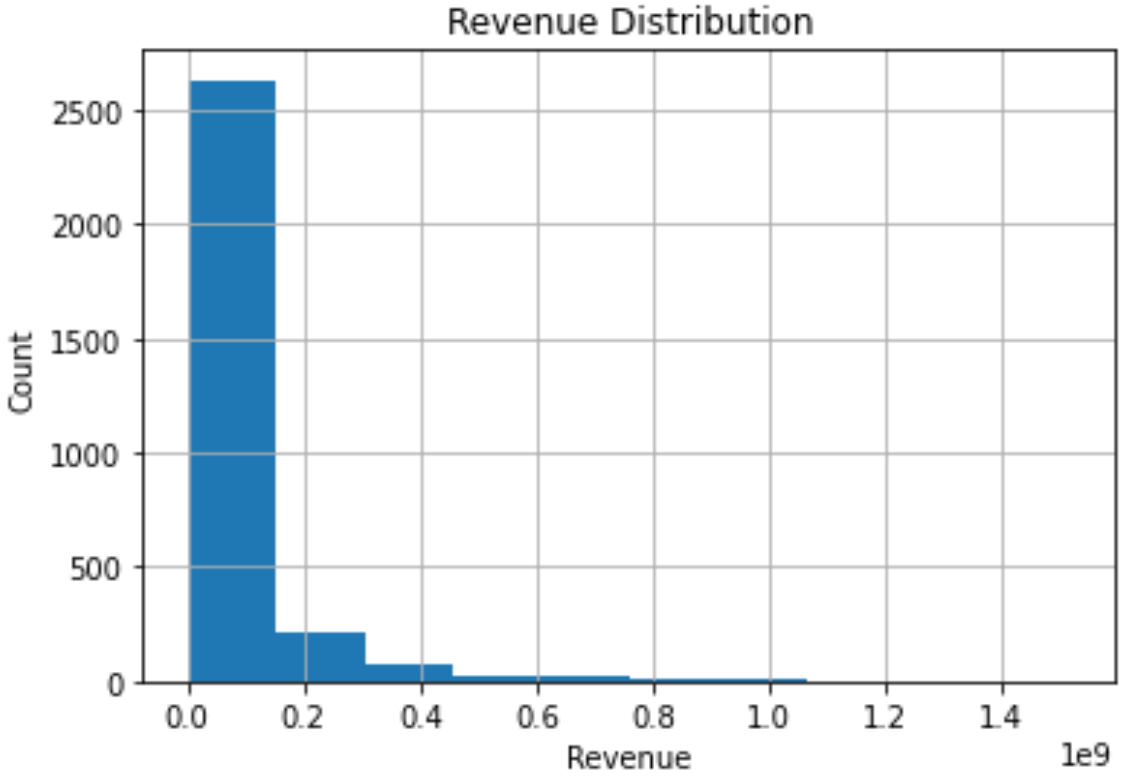
Stakeholders

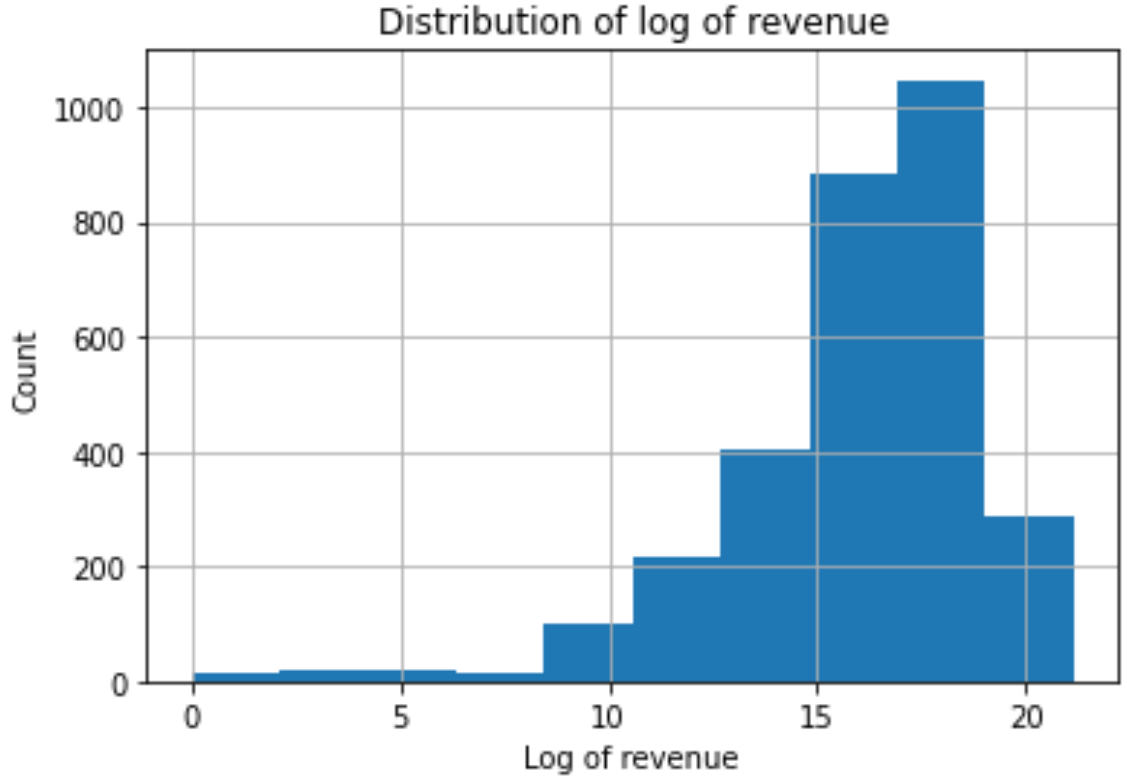


Dataset

- https://www.kaggle.com/competitions/tmdb-box-office-prediction/data

Our target: Revenue



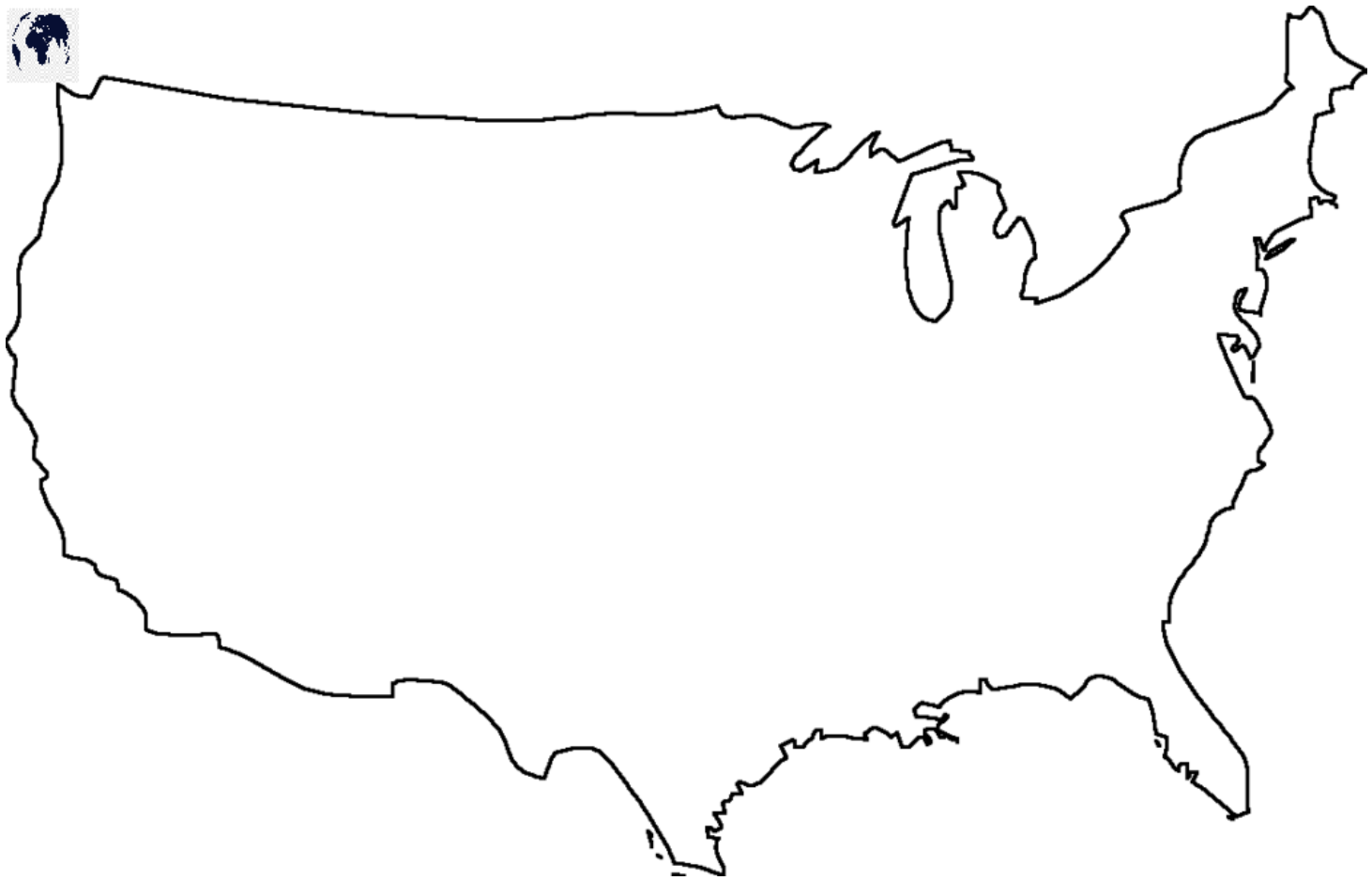


Marketing features:

- Tagline
- Homepage
- Keywords

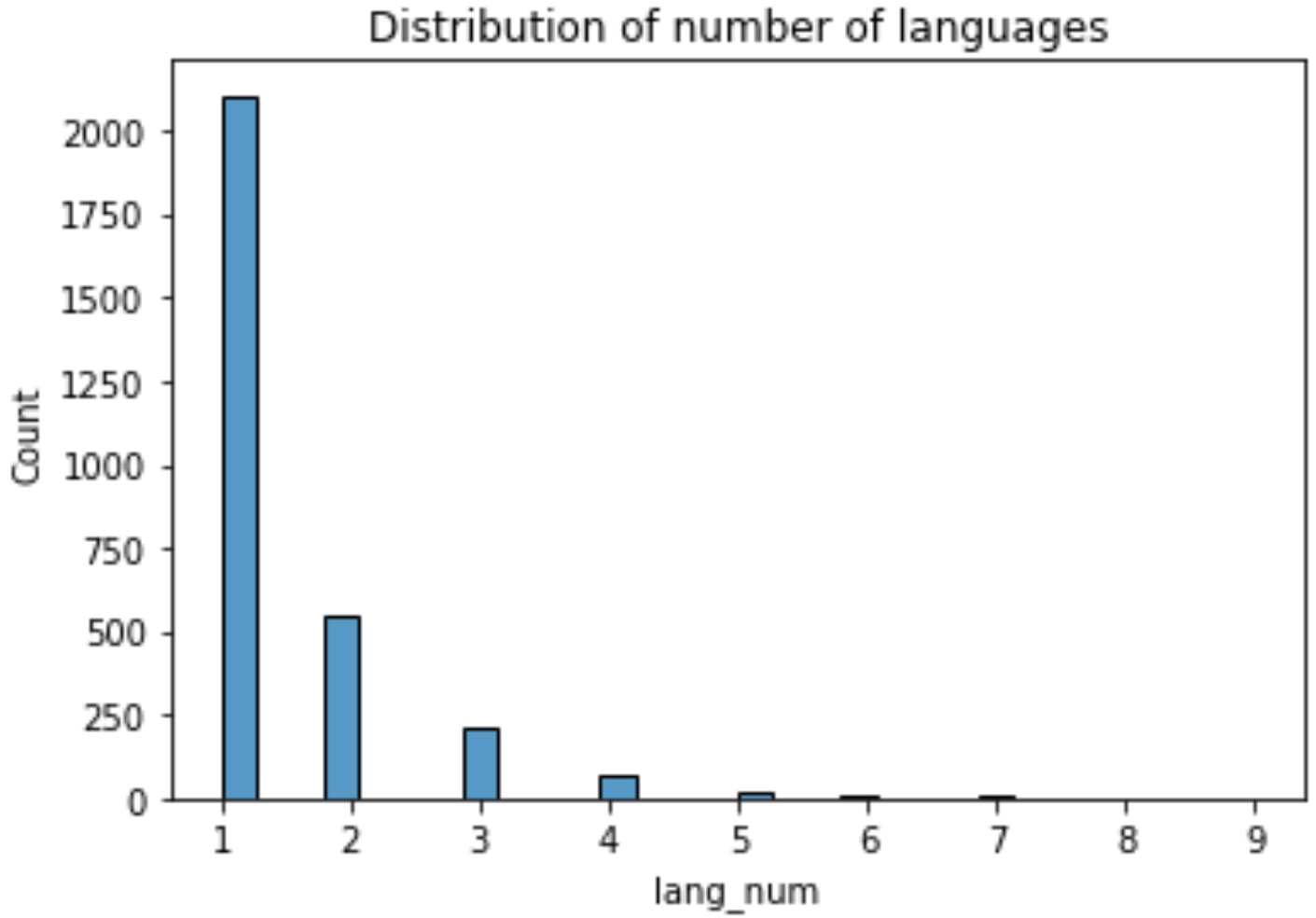
USA-centric productions

- Production Company
- Production Country



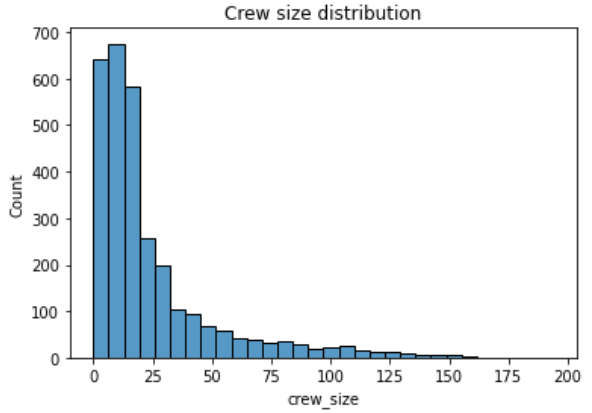
English

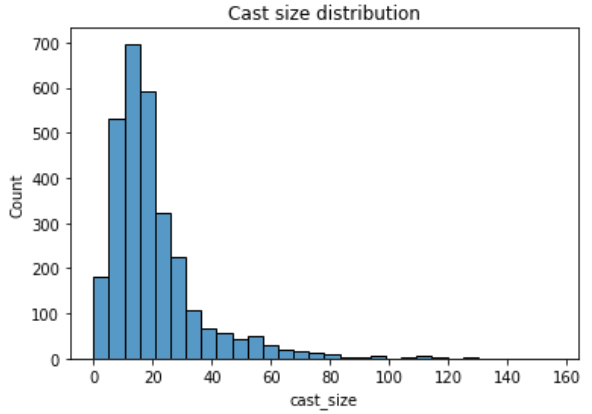
- Is the original language of 86% of films
- Accounts for 60% of all languages



Cast and Crew

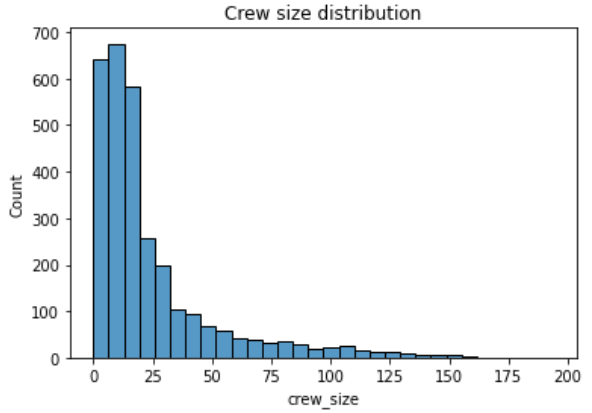
- Size
- Gender
- Prolific personnel

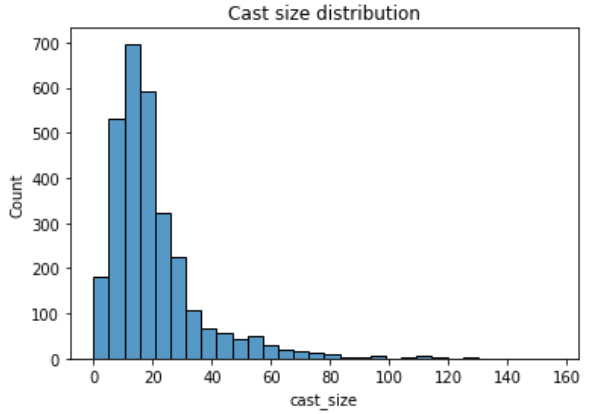




Cast and Crew

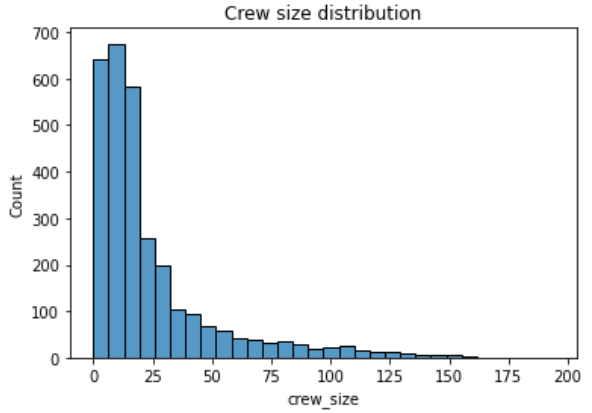
- **Size**
- Gender
- Prolific personnel

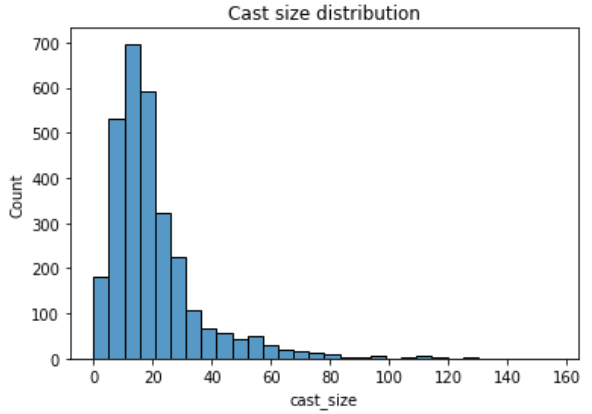




Cast and Crew

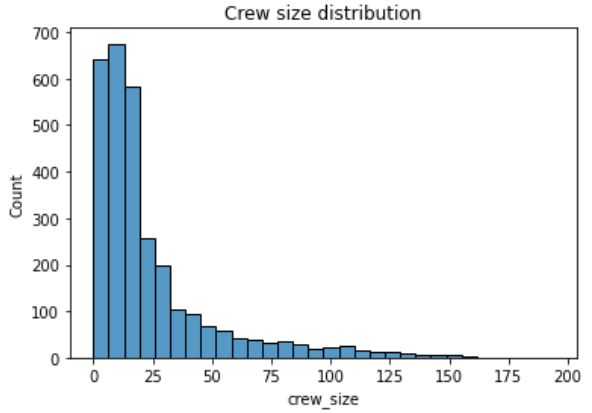
- Size
- **Gender**
- Prolific personnel

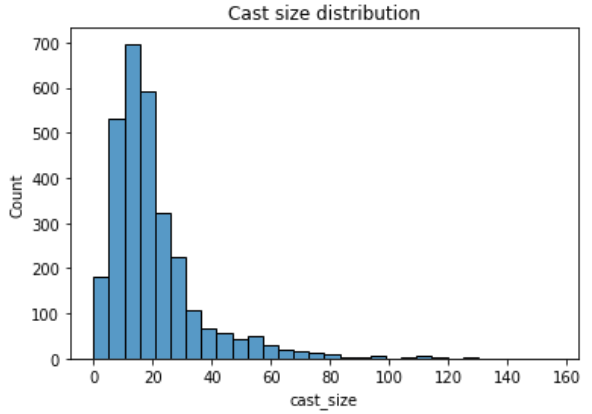




Cast and Crew

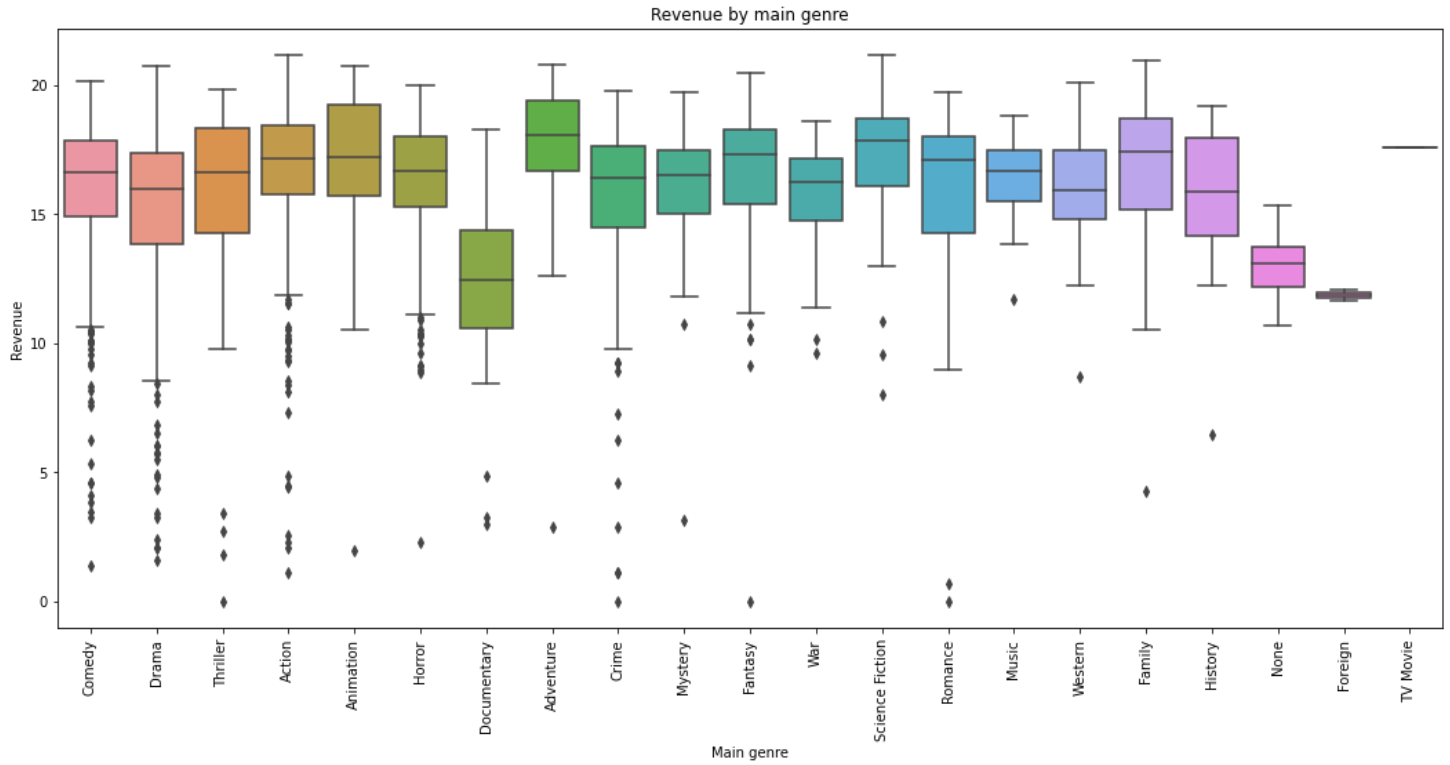
- Size
- Gender
- **Prolific personnel**



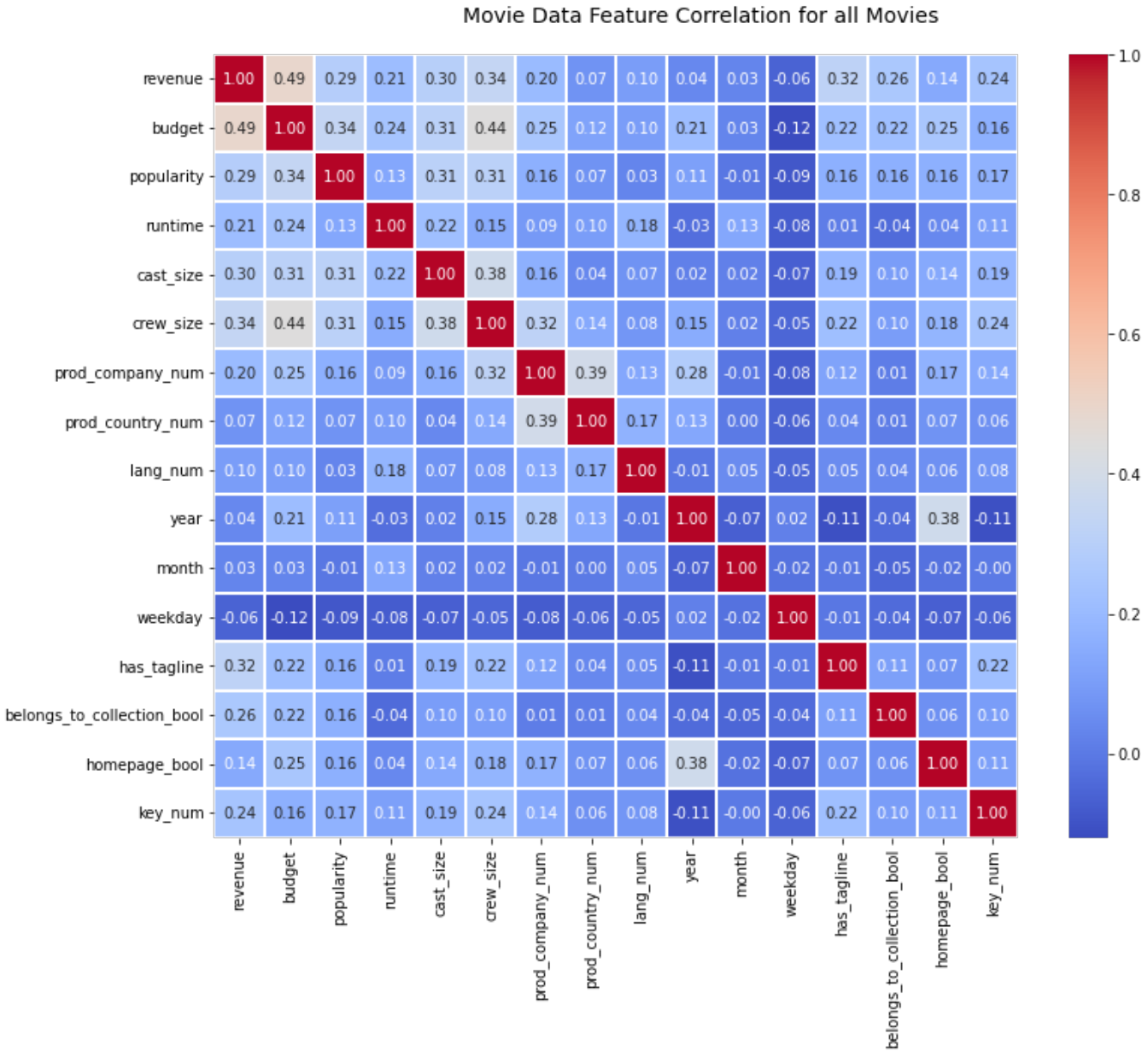


Films are distributed across 21 genres

- More than half of films have 2 or more genres



Numeric Features and Correlations



Scope issues

Americentrism – a problem?

Releases from 1920s to 2018

The Popularity Problem

- Metric tracking positive user interaction with the film on TMDB
- **IS** measured for films in production
- Length of sustained interest is weighted
- Heavily favors released films

23 features to 230 features

Modeling

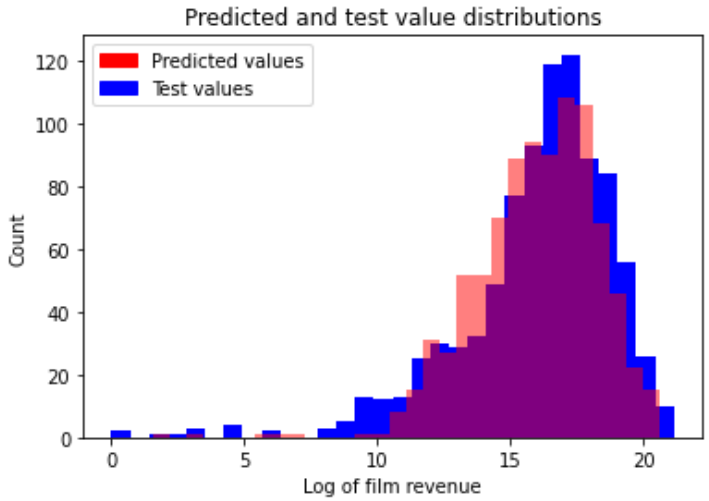
- Linear Regression
- KNeighbors Regression
- Random Forest Regression
- XGBoost Regression

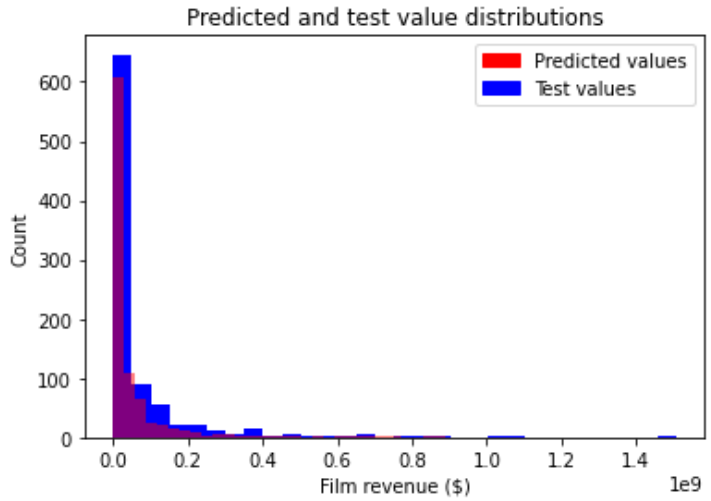
XGBoost Regression Model Has Best Performance

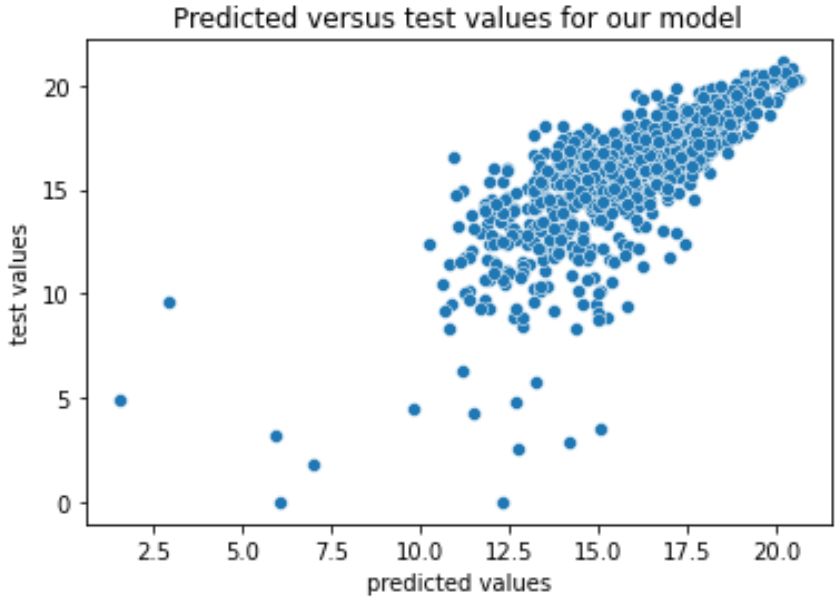
No popularity model performs only slightly worse – **still better than other models**

	train score	test score	RMSE
Linear Regression	0.493	0.402	2.272
KNeighbors Regression	0.524	0.432	2.214
Random Forest Regression	0.813	0.498	2.082
XGBoost Regression	0.803	0.605	1.847
XGBoost no popularity	0.783	0.592	1.878

Model performs better for high revenue films

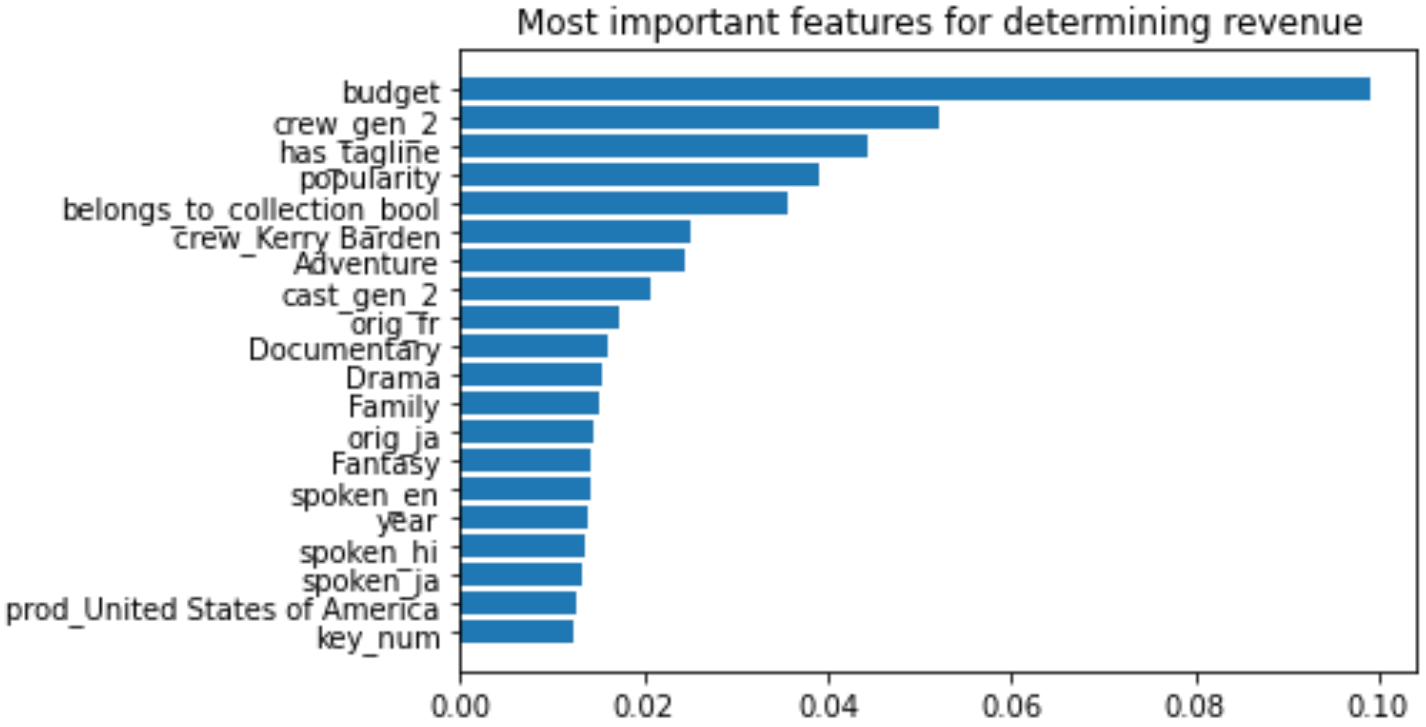






Most important features

- Popularity comes fourth
- Many (103) completely unimportant features



Conclusions

XGB Regression Model No Popularity

	train score	test score	RMSE
Linear Regression	0.493	0.402	2.272
KNeighbors Regression	0.524	0.432	2.214
Random Forest Regression	0.813	0.498	2.082
XGBoost Regression	0.803	0.605	1.847
XGBoost no popularity	0.783	0.592	1.878

Issues and Further Work

The Covid Crash limits predictive capability

Data from covid-heavy years can’t be used for training either

Is the trend away from the box office permanent?

Questions?