# Better Movies Through Data

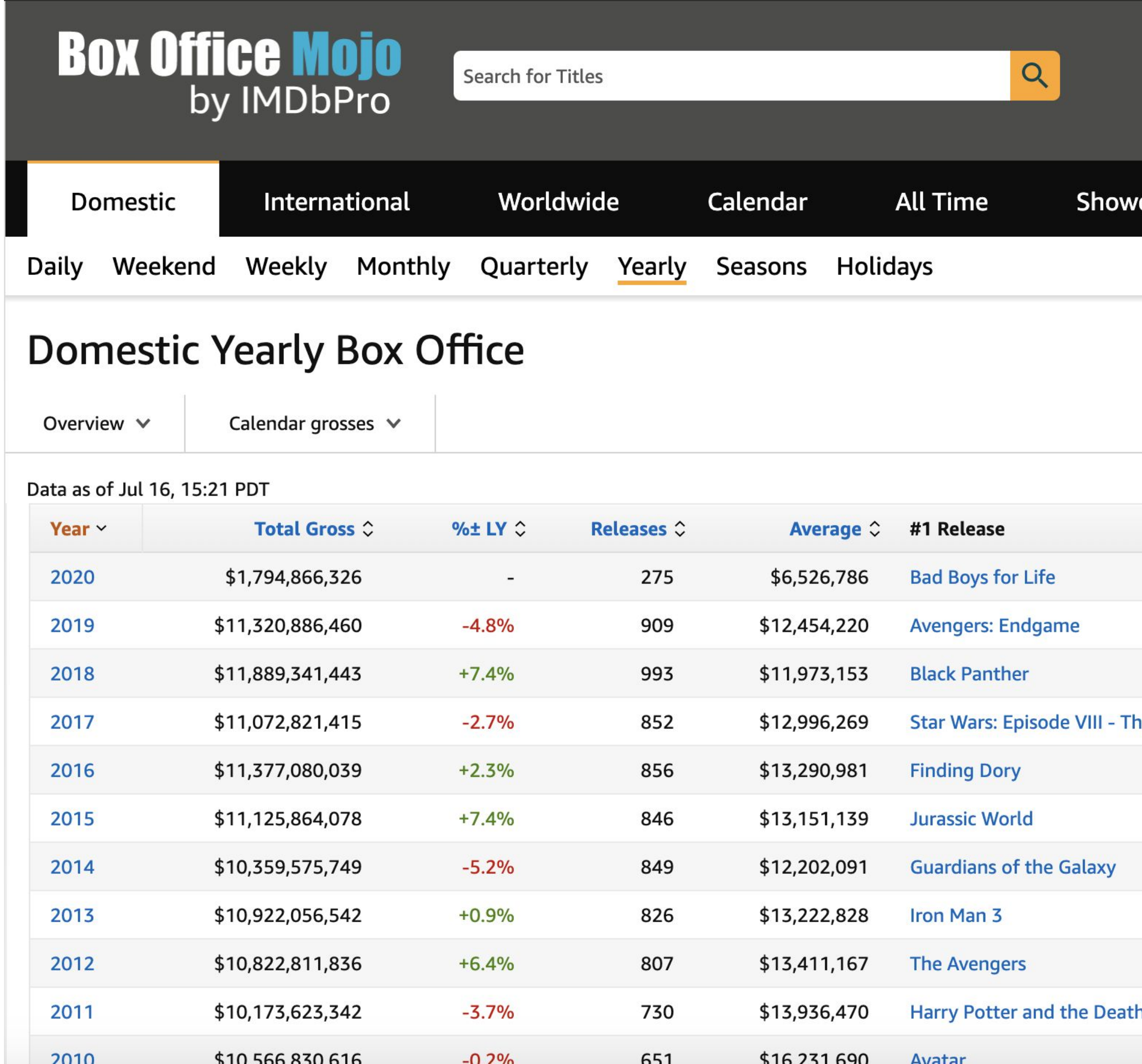
Metis Project II

Stephen DeFerrari

# What's in the Box?

Websites like Box Office Mojo hold the "key" to helping film studios navigate which films to produce, how much to pump into them, and when to release them at what rating in order to ensure domestic box office success

Can we help film studios decide which films to pick up and how they should produce them using simple web scraping and linear regression?



## Domestic Yearly Box Office

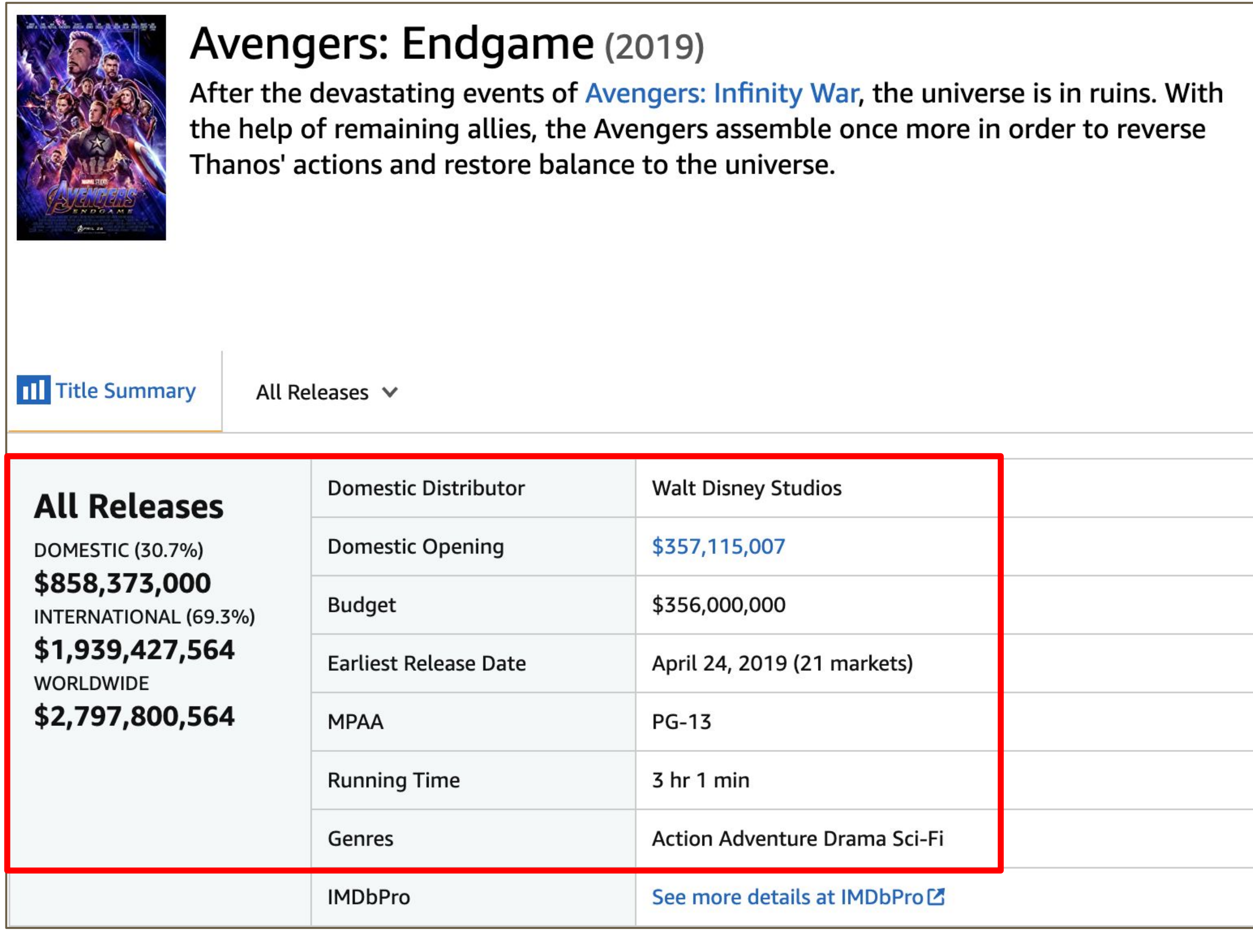
Data as of Jul 16, 15:21 PDT

| Year | Total Gross | %± LY | Releases | Average | #1 Release |
|---|---|---|---|---|---|
| 2020 | $1,794,866,326 | - | 275 | $6,526,786 | Bad Boys for Life |
| 2019 | $11,320,886,460 | -4.8% | 909 | $12,454,220 | Avengers: Endgame |
| 2018 | $11,889,341,443 | +7.4% | 993 | $11,973,153 | Black Panther |
| 2017 | $11,072,821,415 | -2.7% | 852 | $12,996,269 | Star Wars: Episode VIII - Th |
| 2016 | $11,377,080,039 | +2.3% | 856 | $13,290,981 | Finding Dory |
| 2015 | $11,125,864,078 | +7.4% | 846 | $13,151,139 | Jurassic World |
| 2014 | $10,359,575,749 | -5.2% | 849 | $12,202,091 | Guardians of the Galaxy |
| 2013 | $10,922,056,542 | +0.9% | 826 | $13,222,828 | Iron Man 3 |
| 2012 | $10,822,811,836 | +6.4% | 807 | $13,411,167 | The Avengers |
| 2011 | $10,173,623,342 | -3.7% | 730 | $13,936,470 | Harry Potter and the Death |
| 2010 | $10,566,830,616 | -0.2% | 651 | $16,231,690 | Avatar |

# How the Student Stole Data

Going by year from 2019 to 2010 and scraping the top 250 films per year

Scraped all the data in the red box, ultimately tossing international returns



## Avengers: Endgame (2019)

After the devastating events of Avengers: Infinity War, the universe is in ruins. With the help of remaining allies, the Avengers assemble once more in order to reverse Thanos' actions and restore balance to the universe.

### All Releases

DOMESTIC (30.7%)
**$858,373,000**
INTERNATIONAL (69.3%)
**$1,939,427,564**
WORLDWIDE
**$2,797,800,564**

| | |
|---|---|
| Domestic Distributor | Walt Disney Studios |
| Domestic Opening | $357,115,007 |
| Budget | $356,000,000 |
| Earliest Release Date | April 24, 2019 (21 markets) |
| MPAA | PG-13 |
| Running Time | 3 hr 1 min |
| Genres | Action Adventure Drama Sci-Fi |
| IMDbPro | See more details at IMDbPro |

# Special Features

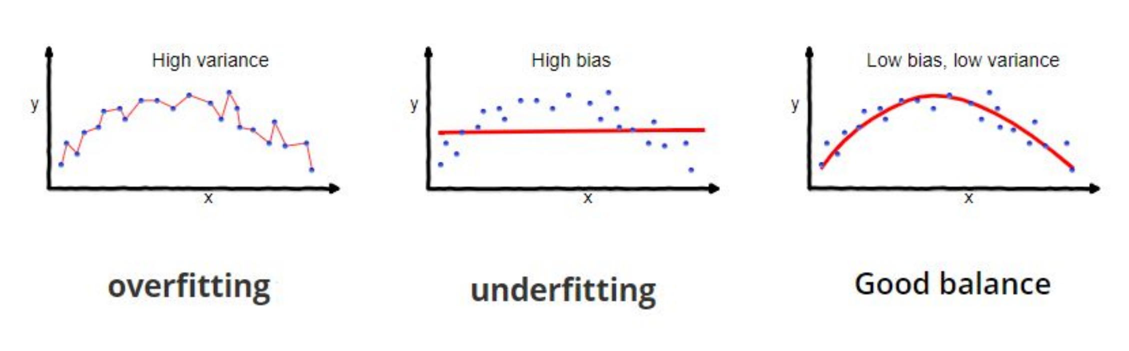
Created a feature checking whether the film was produced by one of the “big six” hollywood studios

Created a feature which checked whether the film’s script had been featured on the Hollywood “Blacklist” of top unmade scripts

# Methodology of Results:
# The Good, The Bad, and the Ugly

Ridge Regression on scaled and polynomial data with low alpha netted me a .60 R^2.

Lasso Regression on scaled and polynomial data with a high alpha netted me a .70 R^2 *

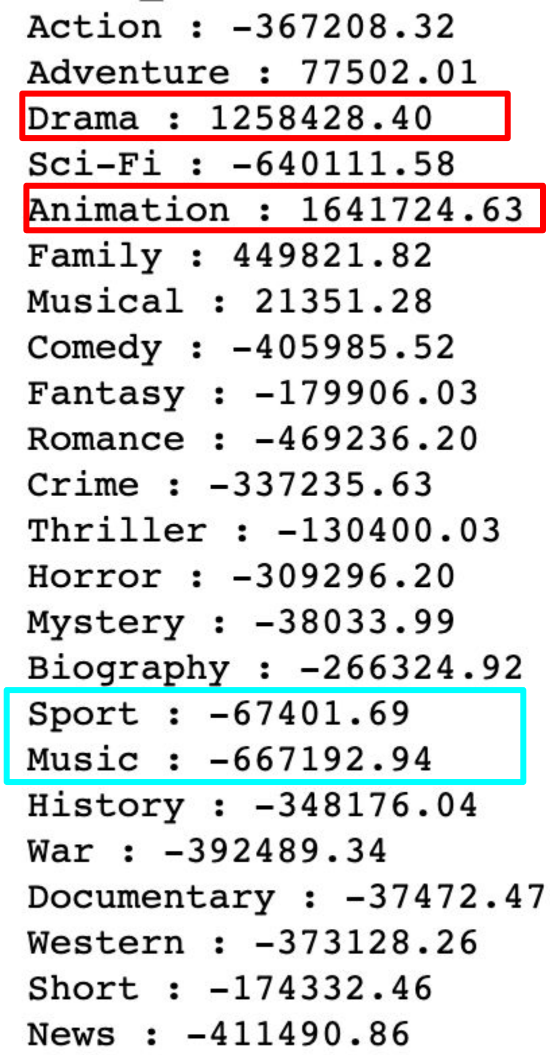


Ridge Regression Results

# Taste of Genre

- Top points for Dramas and Animated films, pointing to two different kinds of audiences you can draw into the movie theater

- On the lower end were sports movies and music movies.
    - Not to be confused with musicals



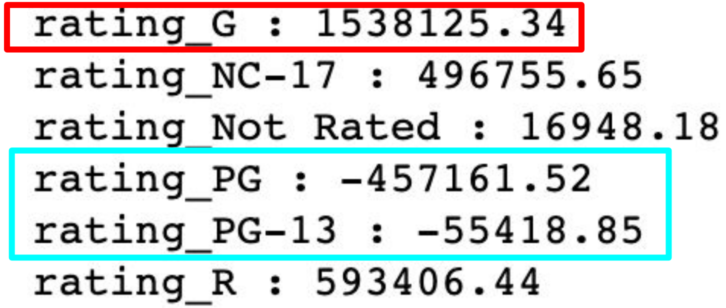
```
Action : -367208.32
Adventure : 77502.01
Drama : 1258428.40
Sci-Fi : -640111.58
Animation : 1641724.63
Family : 449821.82
Musical : 21351.28
Comedy : -405985.52
Fantasy : -179906.03
Romance : -469236.20
Crime : -337235.63
Thriller : -130400.03
Horror : -309296.20
Mystery : -38033.99
Biography : -266324.92
Sport : -67401.69
Music : -667192.94
History : -348176.04
War : -392489.34
Documentary : -37472.47
Western : -373128.26
Short : -174332.46
News : -411490.86
```

# Released in the Month of July: Fierce competition

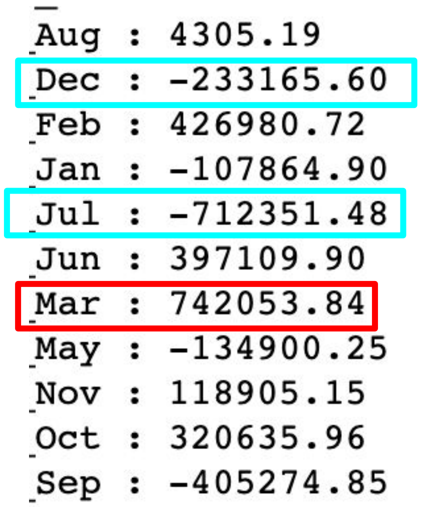
- "General Audience" reigns supreme
  - connected to dominance of animated movies

- March is a secret sweet month while July and December - two months associated with big gains - suffer due to the competition

## Ridge Regression Results



```
rating_G : 1538125.34
rating_NC-17 : 496755.65
rating_Not Rated : 16948.18
rating_PG : -457161.52
rating_PG-13 : -55418.85
rating_R : 593406.44
```



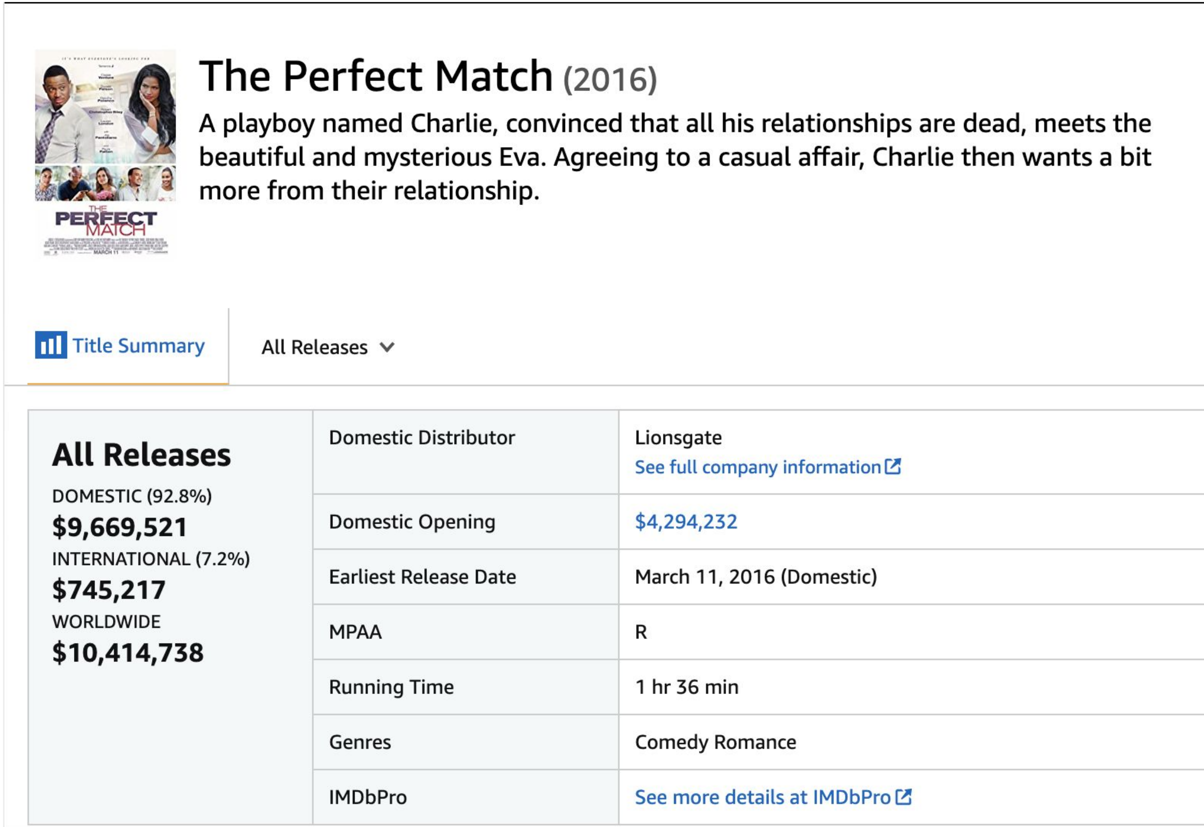
```
_
_Aug : 4305.19
_Dec : -233165.60
_Feb : 426980.72
_Jan : -107864.90
_Jul : -712351.48
_Jun : 397109.90
_Mar : 742053.84
_May : -134900.25
_Nov : 118905.15
_Oct : 320635.96
_Sep : -405274.85
```

# The (Almost) Perfect Match

**Model Prediction:**

**$9,518,385**

**Actual Domestic: $9,669,521**



## The Perfect Match (2016)

A playboy named Charlie, convinced that all his relationships are dead, meets the beautiful and mysterious Eva. Agreeing to a casual affair, Charlie then wants a bit more from their relationship.

**All Releases**

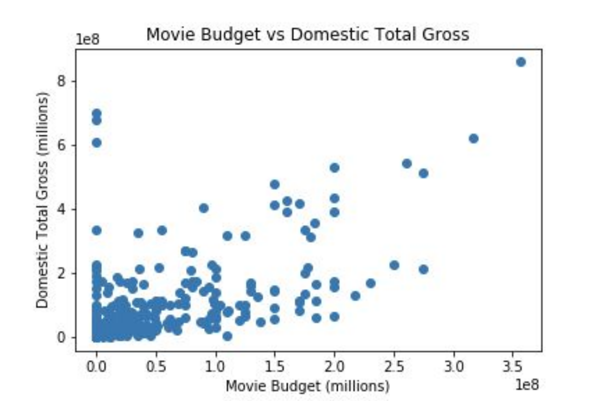
DOMESTIC (92.8%)
**$9,669,521**
INTERNATIONAL (7.2%)
**$745,217**
WORLDWIDE
**$10,414,738**

| | |
|---|---|
| Domestic Distributor | Lionsgate<br>See full company information |
| Domestic Opening | $4,294,232 |
| Earliest Release Date | March 11, 2016 (Domestic) |
| MPAA | R |
| Running Time | 1 hr 36 min |
| Genres | Comedy Romance |
| IMDbPro | See more details at IMDbPro |

# The Limitation Game

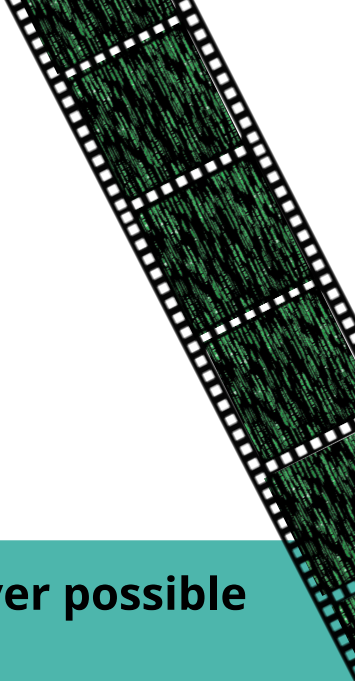
(Where I went wrong)

**“.66” correlation?**



- Over half the movies were missing budgets on their pages
- Dug too deep into years, not wide enough
- Lasso Models would wildly fluctuate scores based on the random state
- There’s still a lot of pieces missing in terms of what makes a movie

# Coming in Better Movies 2: Electric Boogaloo



- Pull From multiple websites and APIs related to films to cover possible budget shortcomings and more missing features

- Engineer ways to gauge star power as well as directors and crew in films

- Change Pulls from 250 films deep to top 100 across more years

# Thank You