DESIGNING THEATER SHOWING SCHEDULES

DATA PREPARATION

Lucas Rozendaal, Jeewon Han, Jay Lawrence, Jillayne Clarke, and Sam Oberly

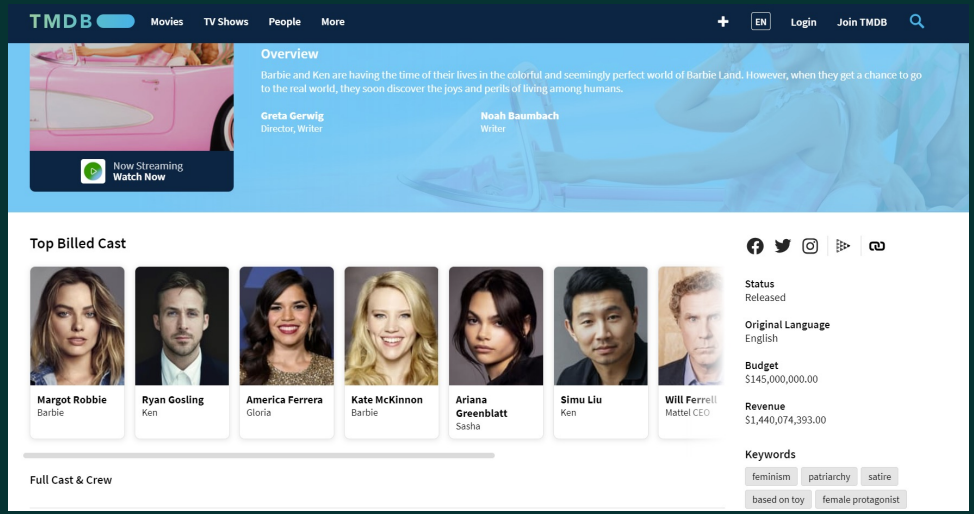
2.1 Initial Data

- Goals:
  - Dataset of Movies
  - Reliable metadata on each movie
  - Usable metrics to compare movies
- Source 1:
  - TMDB
  - Issues: Crowdsourced, Many Missing Datapoints
- Source 2:
  - Box Office Mojo
  - Issues: Needed to scrape
- Source 3:
  - The Numbers
  - Issues: Manual Scraping

2.1 Initial Data – TMDB

- Clean API
- Page Issue
- Missing Data Fields

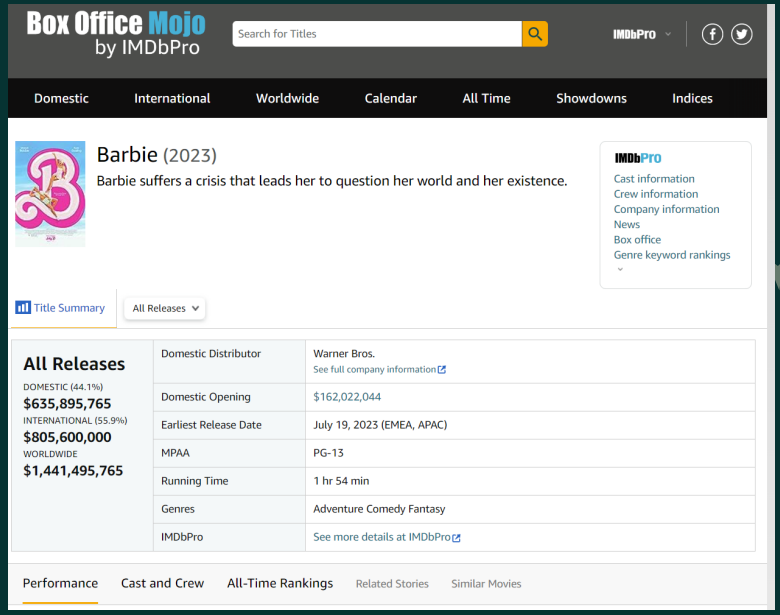
```python
def get_movie_ids_by_year(year):
    pages, results = get_number_pages_and_results_year(year)
    ciel_pages = -(pages //-2)
    data_asc = get_asc_by_year(year, ciel_pages)
    data_desc = get_desc_by_year(year, ciel_pages)

    ids_desc = []
    ids_asc = []
    for p in range(ciel_pages):
        for i in range(len(data_desc[p]["results"])):
            ids_desc.append(data_desc[p]["results"][i]["id"])
        for j in range(len(data_asc[p]["results"])):
            ids_asc.append(data_asc[p]["results"][j]["id"])

    ids = ids_asc + ids_desc
    ids = list(set(ids))
    print(f"We wanted {results}, and we got {len(ids)} for year {year}")

    return ids
```



2.1 Initial Data – Box Office Mojo

- Web Scraped
- Some Missing Data
- Uses IMDb

```python
# Get summary data, title, and revenues
def get_data_by_id(id):
    #read in the URL
    URL = f"https://www.boxofficemojo.com/title/{id}/?ref_=bo_se_r_2"
    page = requests.get(URL)
    soup = BeautifulSoup(page.content, "html5lib")
    #get the data
    title = get_title(soup)
    revenue = get_rev(soup)
    data = get_summary_data(soup)
    data["Title"] = title
    data["Revenue"] = revenue
    #return
    return data
```



2.1 Initial Data – The Numbers

- Specified Ranking Algorithm
- Prevents automatic scraping
- Copy and Pasted ☹

	A	B	C	D	E	F
1	2,023					
2	Rank	Name	Star Score	Movies	Average Billing	
3	1	Margot Robbie	323	5	4.80	
4	2	Dave Bautista	291	4	4.80	
5	3	Michelle Rodriguez	263	3	3.70	
6	4	Hiroyuki Sanada	246	3	7.70	
7	5	Zoe Saldana	238	3	3.30	
8	6	David Harbour	231	3	2	
9	7	Tom Holland	228	3	1	

Highest Grossing Stars of 2023 at the Domestic Box Office

See also: Top Domestic 2018 Stars - Top Domestic 2019 Stars - Top Domestic 2020 Stars - Top Domestic 2021 Stars - Top International 2023 Stars - Top Worldwide 2023 Stars
This list shows the highest grossing stars of 2023 based on the domestic box office of the movies they had a leading role in in 2023 and the two preceeding years (voice-only roles excluded).

The Star Score in this list represent points assigned to each of the leading stars of top 100 (based on domestic box office) movies in the current year and two preceding years. For appearing in the number one movie in a year a star gets 100 points, the number two movie 99 points and so on. This rewards actors for appearing in a number of hit films over the course of three years more than starring in just one monster hit over the same time period.

Top Domestic 2022 Stars

1

Margot Robbie

Best known as a **Leading Actress** based on credits in that role in 13 films.

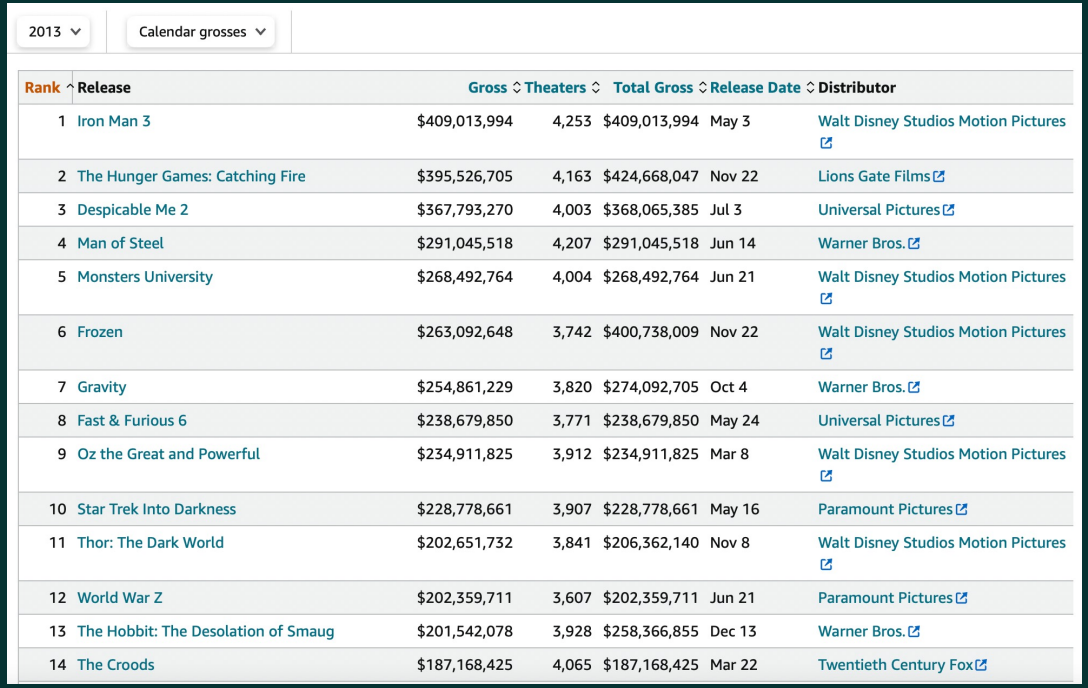
Top films contributed to this record: Barbie ($635,913,366 in 2023), The Suicide Squad ($55,817,425 in 2021), Asteroid City ($28,153,025 in 2023)

Full Information...

2.4 Data Quality

	98357	56000000	[53, 80, 18]	['Thriller', 'Crime', 'Dra	tt1235522	[8532, 10405	['1984 Privat	1/18/13	34737199	109	Broken City
1443	399217	0	[18, 10749]	['Drama', 'Romance']	tt3467914	[]	[]	7/7/17	751345	117	Tommy's Honour
2182	535292	33000000	[80, 28, 18]	['Crime', 'Action', 'Dra	tt8688634	[106544, 362	['AGBO', 'MV	10/24/19	49939757	99	21 Bridges
37	233364	0	[35, 10749]	['Comedy', 'Romance']	tt2769454	[4676]	['Star Cinem	3/12/13	10000000	100	Must Be... Love
2611	484297	0	[18, 35]	['Drama', 'Comedy']	tt5805768	[17090, 9204	['FJ Productic	12/24/20	28657	85	Abe
2552	443791	50000000	[27, 878, 28,	['Horror', 'Science Fict	tt5774060	[7076, 25, 22	['Chernin Ent	1/8/20	40882928	95	Underwater
347	230179	8500000	[28, 12, 53]	['Action', 'Adventure',	tt2088003	[62407, 624(	['Bavaria Filr	9/5/14	7500000	90	Big Game
316	226354	0	[18, 10751]	['Drama', 'Family']	tt2739338	[39281, 1028	['EchoLight S	11/22/13	2476775	100	The Christmas Candle
3201	942624	0	[18]	['Drama']	tt8110900	[184416, 54(	['UP2U Films	9/16/22	1445862	127	Running the Bases
3395	1097185	5	[53]	['Thriller']		[]	[]	4/14/23	6	4	The Soulsman
3144	934455	0	[16, 35]	['Animation', 'Comedy']		[]	[]	2/3/22	1	5	Fart: The Movie Pilot
1456	432942	0	[18]	['Drama']	tt6057032	[4560, 9079(	['Tunnel Post	8/18/17	250130	94	Gook
1190	376290	13000000	[18, 80]	['Drama', 'Crime']	tt4540710	[7493, 4818E	['FilmNation	11/25/16	9101546	133	Miss Sloane
1781	463821	42000000	[14, 10751, 3	['Fantasy', 'Family', 'C	tt2119543	[56, 34982, 7	['Amblin Ent	9/15/18	131523093	105	The House with a Clock in
2272	608793	450	[878, 28, 35]	['Science Fiction', 'Action', 'Comedy		[116640]	['City Lights I	7/1/19	450	3	OVERTIME
2198	635100	0	[99, 10402]	['Documentary', 'Musi	tt11041584	[3447]	['IMAX']	10/25/19	1082629	31	Jesus Is King
2621	747968	50	[16, 14, 1074	['Animation', 'Fantasy', 'Romance']		[]	[]	10/23/20	20	3	Venus
418	238603	13000000	[10751, 12, 8	['Family', 'Adventure',	tt2183034	[10426, 2]	['Panay Film	6/14/14	45300000	89	Earth to Echo
1932	608443	125	[]	[]	tt8571866	[]	[]	3/16/18	40	105	Ne Var?
3150	804095	40000000	[18]	['Drama']	tt14208870	[56, 190543]	['Amblin Ent	11/11/22	45598614	151	The Fabelmans
2690	458576	60000000	[28, 14, 12]	['Action', 'Fantasy', 'Ad	tt6475714	[7220, 47, 24	['CAPCOM', '	12/3/20	42145959	104	Monster Hunter
447	241771	7000000	[10749, 18]	['Romance', 'Drama']	tt3125324	[42321, 7295	['Black Enter	11/14/14	14618727	116	Beyond the Lights
466	177047	2000000	[18]	['Drama']	tt1464191	[13238, 1323	['Unified Pict	5/2/14	60048	91	Decoding Annie Parker
919	253626	0	[18]	['Drama']	tt3297330	[13923, 1947	['Sobini Film	4/9/15	316472	104	Good Kill
2655	619991	1000	[]	[]	tt10651098	[]	[]	11/1/20	1000	0	Pendekar Cyborg
2457	988476	5000	[27]	['Horror']		[198586]	['FruitBird Fi	1/3/20	500	11	Willowpede
1012	262841	125000000	[28, 35, 878]	['Action', 'Comedy', 'Sc	tt3095734	[2348, 4, 102	['Nickelodeo	12/21/16	64493915	104	Monster Trucks
707	263109	25000000	[10751, 16, 3	['Family', 'Animation',	tt2872750	[297, 20664,	['Aardman',	2/5/15	106209378	85	Shaun the Sheep Movie
608	419125	2000	[18, 9648]	['Drama', 'Mystery']	tt3731706	[82015]	['Metamora	5/16/14	500	14	Last Day With Lizzy
3023	986717	25593	[18, 53, 9648	['Drama', 'Thriller', 'Mystery']		[176963]	['Mani Sriniv	6/2/22	5000	135	Flight Maayam
535	282296	5000000	[10749, 35, 1	['Romance', 'Comedy',	tt2908856	[35758, 9122	['Deux Cheva	9/9/14	7527232	107	My Old Lady
3077	992432	120000	[]	[]	tt21136616	[]	[]	6/25/22	365000	50	The Secret Of Palestine
3426	872585	100000000	[18, 36]	['Drama', 'History']	tt15398776	[9996, 33, 5(	['Syncopy', 'U	7/19/23	925832975	181	Oppenheimer
611	157099	5000000	[18, 35]	['Drama', 'Comedy']	tt1609479	[25837, 2583	['Ealing Metr	3/14/14	75143	91	Better Living Through Cher
2787	337404	200000000	[35, 80]	['Comedy', 'Crime']	tt3228774	[2, 3538, 252	['Walt Disne	5/26/21	233503234	134	Cruella

2.4 Data Quality



2013 | Calendar grosses

Rank	Release	Gross	Theaters	Total Gross	Release Date	Distributor
1	Iron Man 3	$409,013,994	4,253	$409,013,994	May 3	Walt Disney Studios Motion Pictures
2	The Hunger Games: Catching Fire	$395,526,705	4,163	$424,668,047	Nov 22	Lions Gate Films
3	Despicable Me 2	$367,793,270	4,003	$368,065,385	Jul 3	Universal Pictures
4	Man of Steel	$291,045,518	4,207	$291,045,518	Jun 14	Warner Bros.
5	Monsters University	$268,492,764	4,004	$268,492,764	Jun 21	Walt Disney Studios Motion Pictures
6	Frozen	$263,092,648	3,742	$400,738,009	Nov 22	Walt Disney Studios Motion Pictures
7	Gravity	$254,861,229	3,820	$274,092,705	Oct 4	Warner Bros.
8	Fast & Furious 6	$238,679,850	3,771	$238,679,850	May 24	Universal Pictures
9	Oz the Great and Powerful	$234,911,825	3,912	$234,911,825	Mar 8	Walt Disney Studios Motion Pictures
10	Star Trek Into Darkness	$228,778,661	3,907	$228,778,661	May 16	Paramount Pictures
11	Thor: The Dark World	$202,651,732	3,841	$206,362,140	Nov 8	Walt Disney Studios Motion Pictures
12	World War Z	$202,359,711	3,607	$202,359,711	Jun 21	Paramount Pictures
13	The Hobbit: The Desolation of Smaug	$201,542,078	3,928	$258,366,855	Dec 13	Warner Bros.
14	The Croods	$187,168,425	4,065	$187,168,425	Mar 22	Twentieth Century Fox

2.4 Data Quality – Identified Quality Issues

1. Timeout Issue on API Calls in 2022
    - Easy to Resolve, Resolved
2. 12 Individual Films Not in the Dataset
    - Theoretically Easy to Resolve, Currently Unresolved
3. Foreign Films in Dataset/Release Date Issues
    - Difficult to Resolve, Currently Unresolved

3.1 Select Data – Timeframe

Timeframe	Description	Considerations
1888–2023	All years available on TMDB	– Computing Power – Predictive Power (inflation, technological innovations)
2000–2023	Years with comparable socioeconomic conditions	– Poor Box Office Mojo Data Quality pre–2013
2013–2023	Years for final dataset	– Created out of necessity; smaller subset of data than would have liked

3.4 Integrate Data – Original Methodology

Step 1

Include movies with either a TMDB or Box Office Mojo budget

Step 2

AND also TMDB revenue or Box Office Mojo worldwide revenue

Decision Rule

If conflicting information, Box Office Mojo > TMDB

3.4 Integrate Data – Original Methodology

TMDB Revenue

Either worldwide OR domestic revenue depending on what's available

VS

Box Office Mojo Revenue

3 reported figures:

1. **Worldwide**
2. Domestic
3. International

3.4 Integrate Data – Original Methodology

- We assumed Box Office Mojo > TMDB, but had 118 movies with TMDB revenue but *no* Box Office Mojo revenues
  - Opposite of what we expected
  - Why?
- Most were miscategorized movies on TMDB – non-theatrical releases categorized as theatrical
  - Default on TMDB is 'theatrical'
- **Solution**: only 118 movies, so drop

3.4 Integrate Data – Revised Methodology

Step 1

Include movies with either a TMDB or Box Office Mojo budget

Step 2

AND only Box Office Mojo domestic revenue

Decision Rule

If conflicting information, Box Office Mojo > TMDB

3.3 Construct Data

- The Numbers Star Score for Actors
  - Based on Top 100 Highest Domestic Grossing Films Each Year
  - Sum of points for current year and 2 years prior
- Weighting Functions
- Production Company and Director Metrics
- Features from Release Date

Star Score Example

Margot Robbie (2023) - 315

78	44	42	100	51
The Suicide Squad (2021)	**Babylon (2022)**	**Amsterdam (2022)**	**Barbie (2023)**	**Asteroid City (2023)**
Ranked 23rd	Ranked 57th	Ranked 59th	Ranked 1st	Ranked 50th

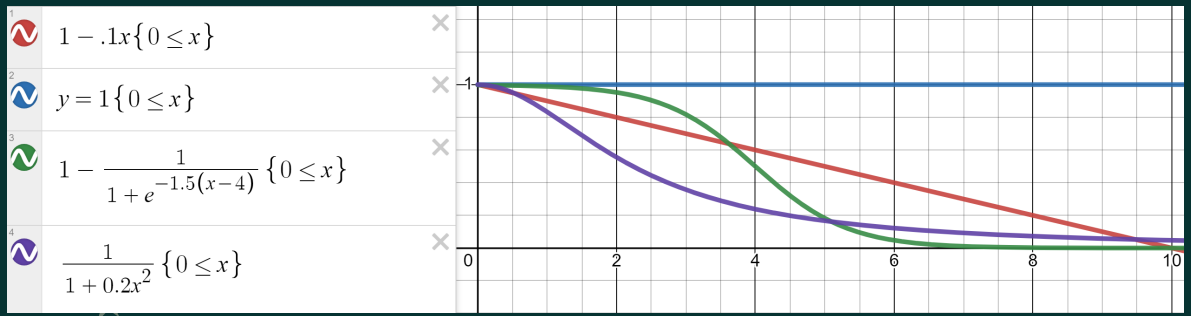
Weight Functions

Criteria:

- More stars means higher star power
- Diminishing returns
- Small casts not penalized

Options:

- Uniform Weight
- Linear Weight
- Log Weight
- Exponential Weight



More Engineered Features

Director and Production Scores

- Director: No weight
- Production: No weight

Release Date

- Release Year
- Season
- Holiday

2.2 Describe Data

Categorical	19
Numerical	18
Engineered	7 (from Numerical)

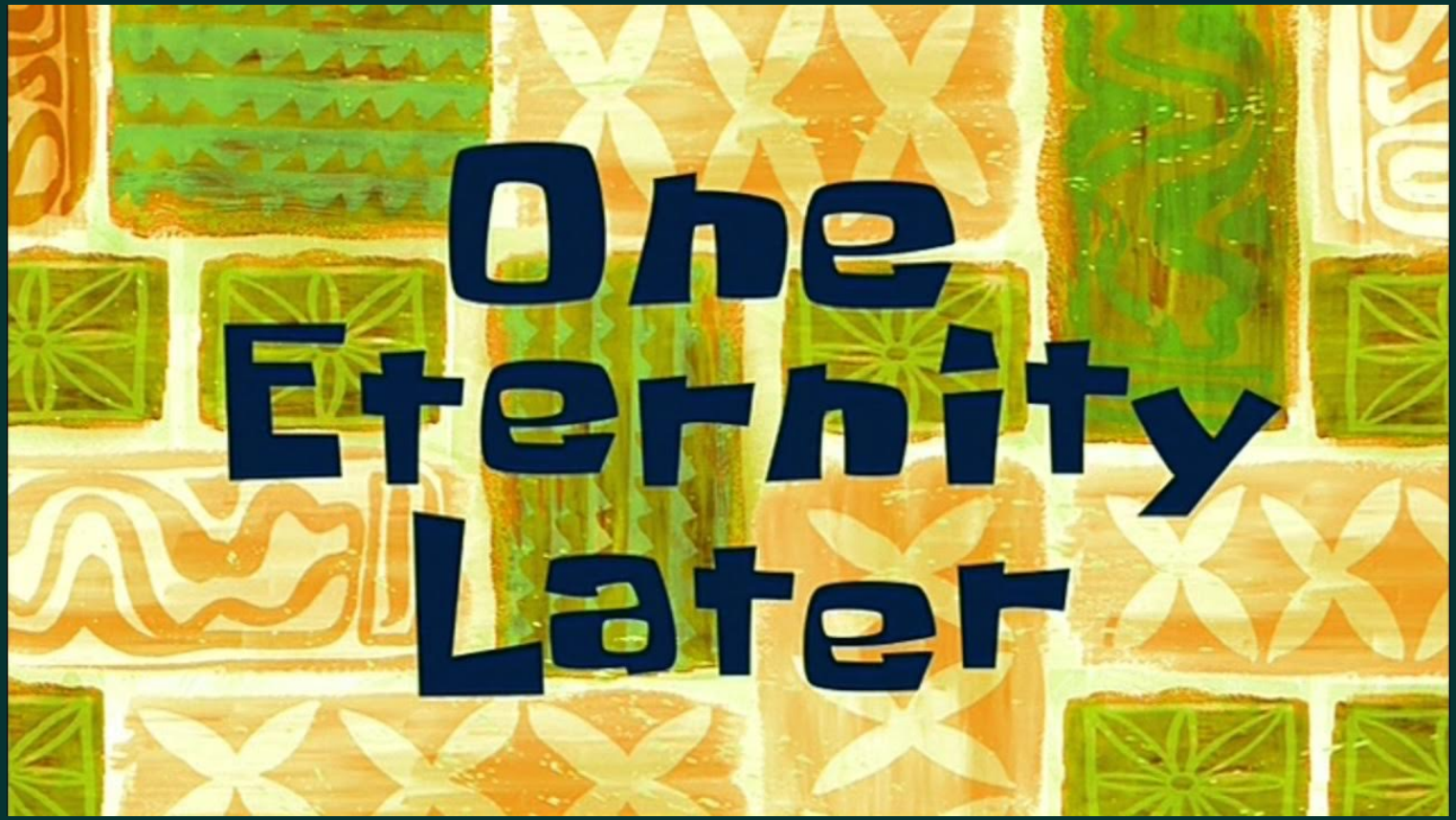
Variable Name	Type	Notes
Unnamed: 0	Categorical	Index
Running Time	Numerical	hrs mins
Genres	Categorical	
IMDB Title	Categorical	
MPAA	Categorical	Maturity Rating
Domestic Distributor	Categorical	
Domestic Opening	Categorical	
Earliest Release Date	Numerical	

3.5 Format Data

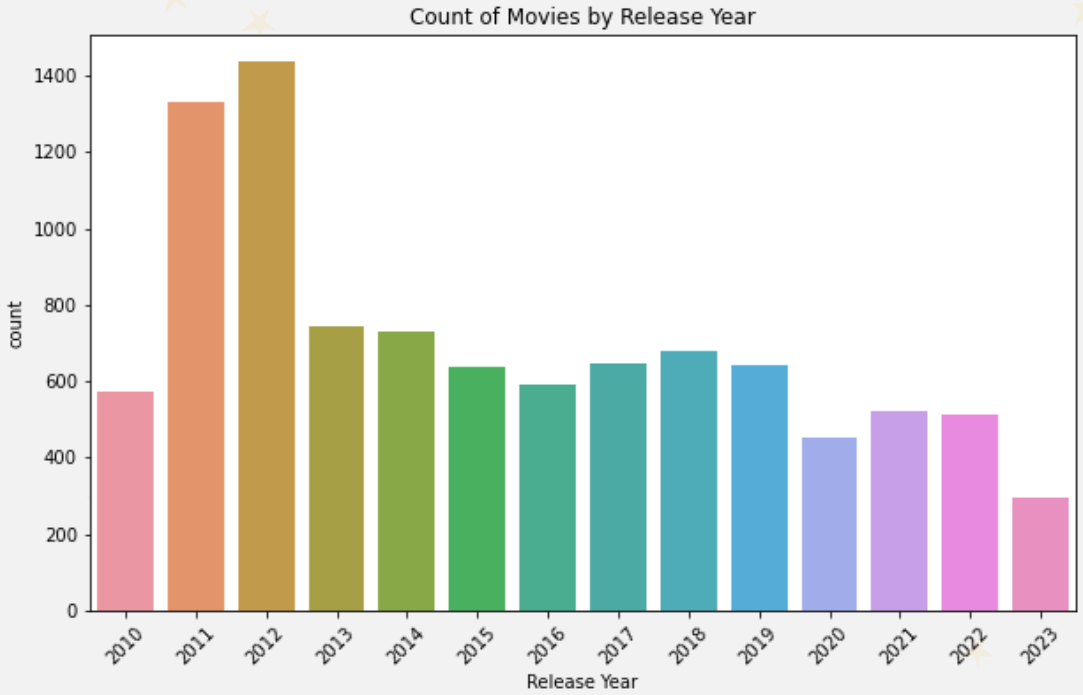
Rename Variables

Remove Redundant Variables

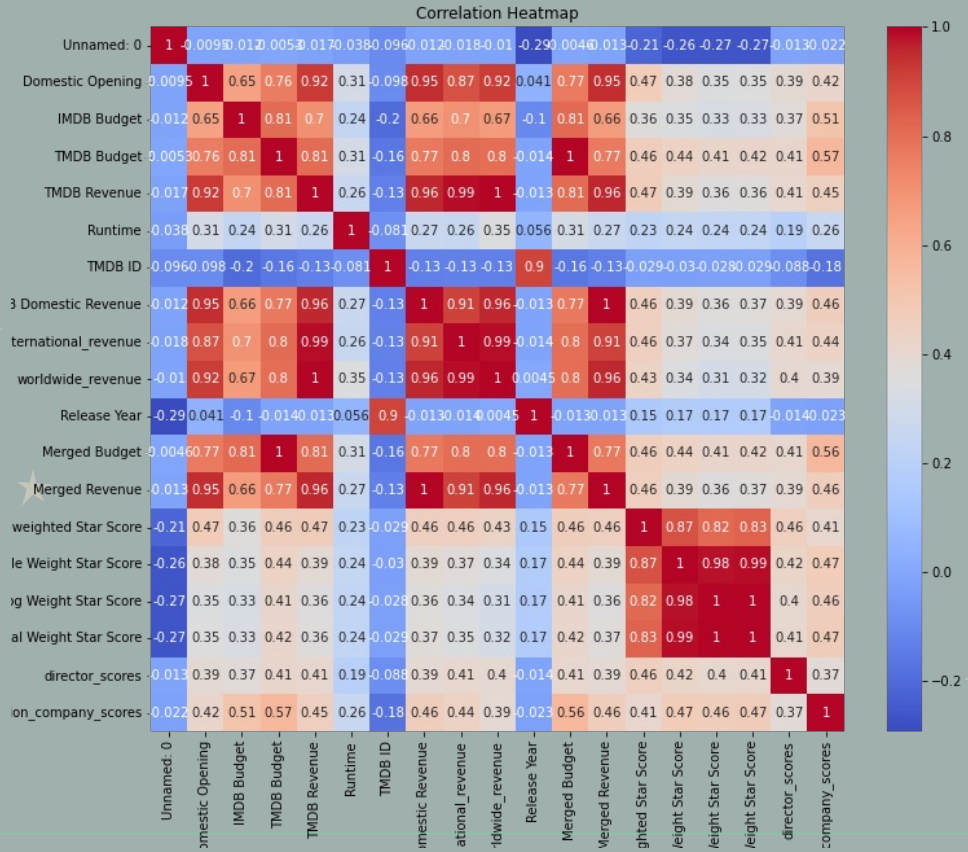
Drop Categorical Variables


One
Eternity
Later

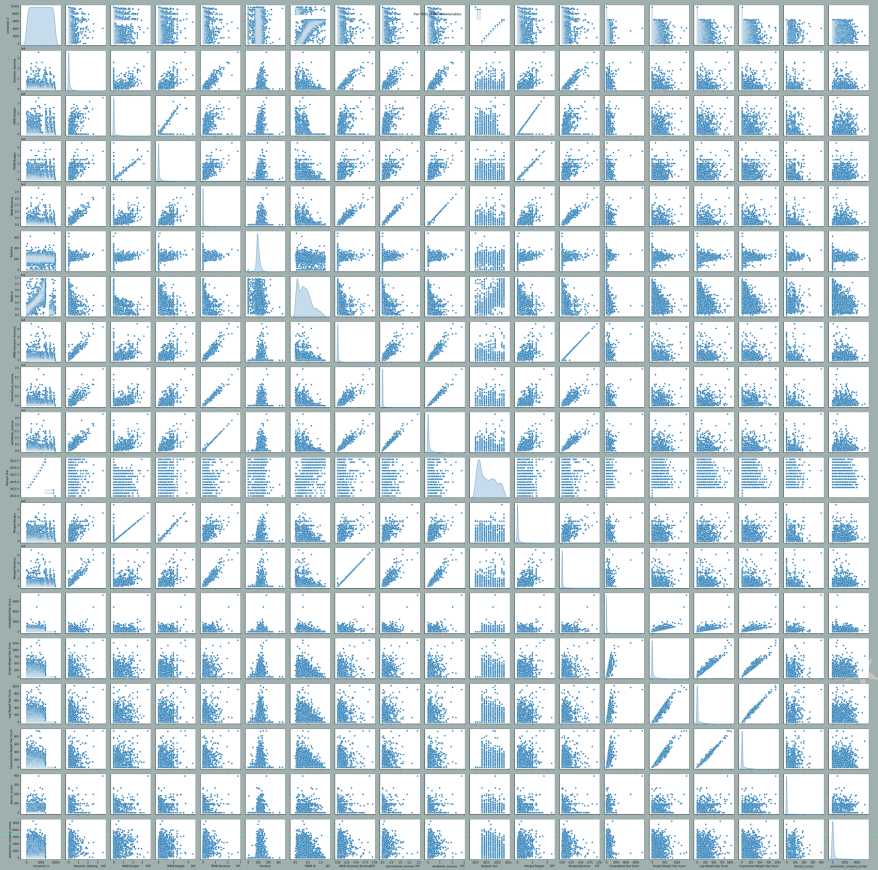
Our Final Dataset!



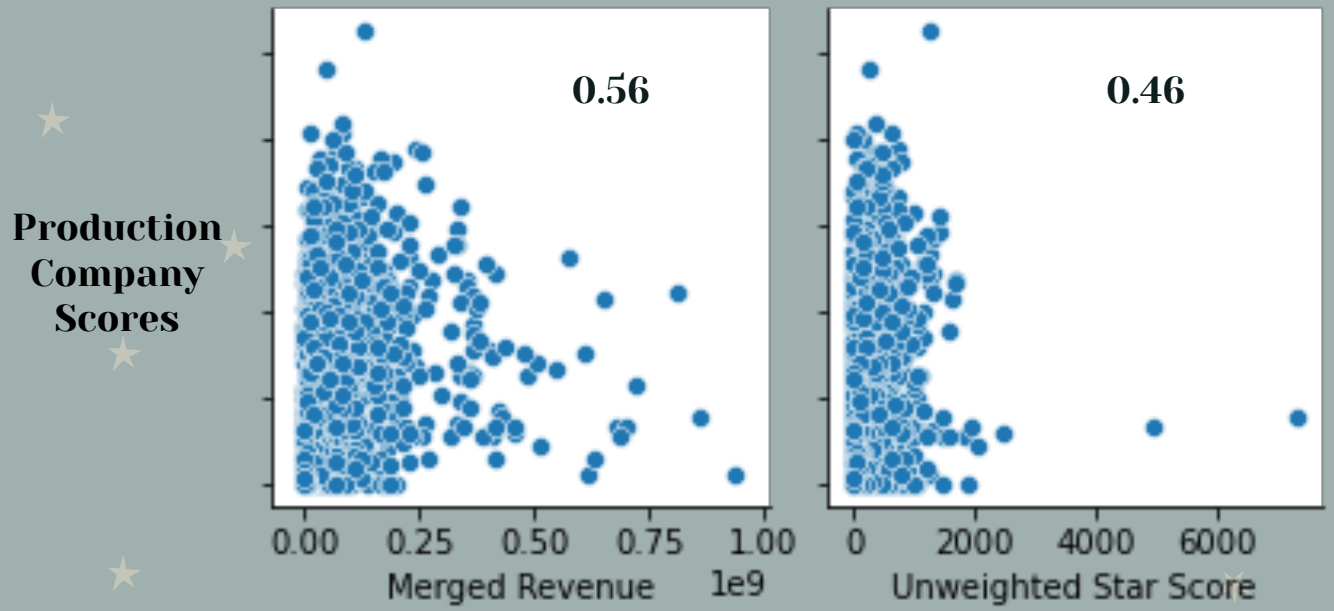
2.3 Explore Data- Heatmap



2.3 Explore Data- Correlation Matrix



Ex: Correlation Matrices: Production Company Scores



Ex: Correlation Matrices: Star Scores