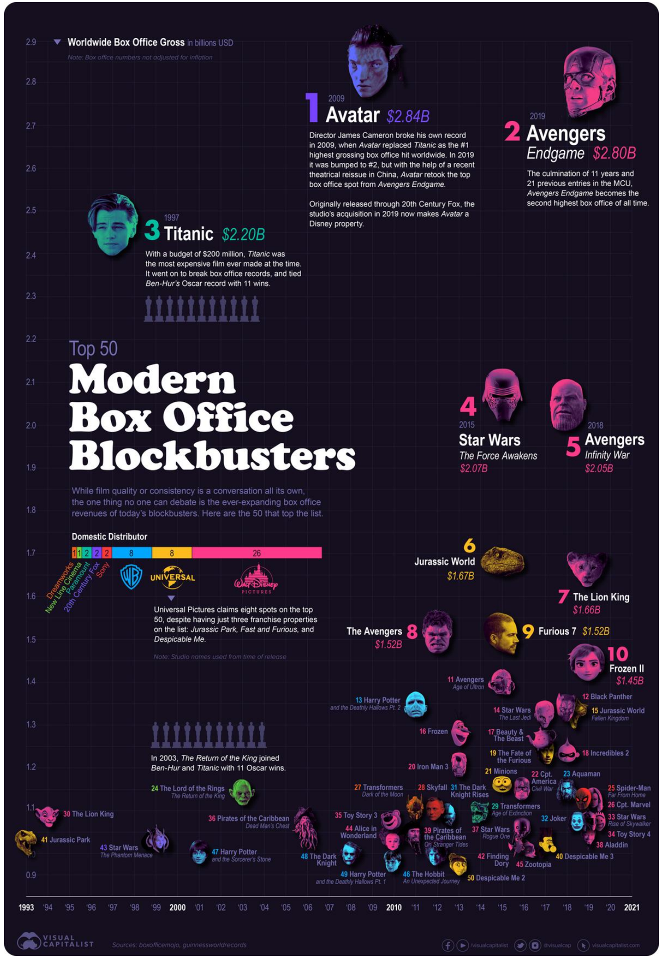
Build a Data Story Final Project

By: Dorothy Kunth

As a data science consultant for a large production company, Spectacular Studios, the tasks involve defining a problem that is compelling enough for the executive team to warrant taking action and to develop the storyline of what can be expected of the data to tell from the analysis.

Dataset - a movies metadata from ***The Movies Dataset*** available on Kaggle
https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

The factors that influence the people’s decision to see a movie



Top 50 **Modern Box Office Blockbusters**

While film quality or consistency is a conversation all its own, the one thing no one can debate is the ever-expanding box office revenues of today’s blockbusters. Here are the 50 that top the list.

Note: Box office numbers not adjusted for inflation

Director James Cameron broke his own record in 2009, when *Avatar* replaced *Titanic* as the #1 highest grossing box office hit worldwide. In 2019 it was bumped to #2, but with the help of a recent theatrical reissue in China, *Avatar* retook the top box office spot from *Avengers Endgame*.

Originally released through 20th Century Fox, the studio’s acquisition in 2019 now makes *Avatar* a Disney property.

The culmination of 11 years and 21 previous entries in the MCU, *Avengers Endgame* becomes the second highest box office of all time.

With a budget of $200 million, *Titanic* was the most expensive film ever made at the time. It went on to break box office records, and tied *Ben-Hur’s* Oscar record with 11 wins.

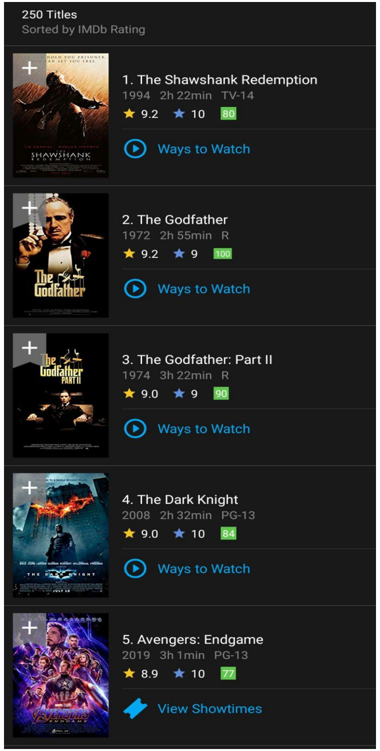
Universal Pictures claims eight spots on the top 50, despite having just three franchise properties on the list: *Jurassic Park, Fast and Furious,* and *Despicable Me.*

Note: Studio names used from time of release

In 2003, *The Return of the King* joined *Ben-Hur* and *Titanic* with 11 Oscar wins.

Sources: boxofficemojo, guinnessworldrecords

Executive Summary



250 Titles
Sorted by IMDb Rating

1. The Shawshank Redemption
1994 2h 22min TV-14
★ 9.2 ★ 10 80
Ways to Watch

2. The Godfather
1972 2h 55min R
★ 9.2 ★ 9 100
Ways to Watch

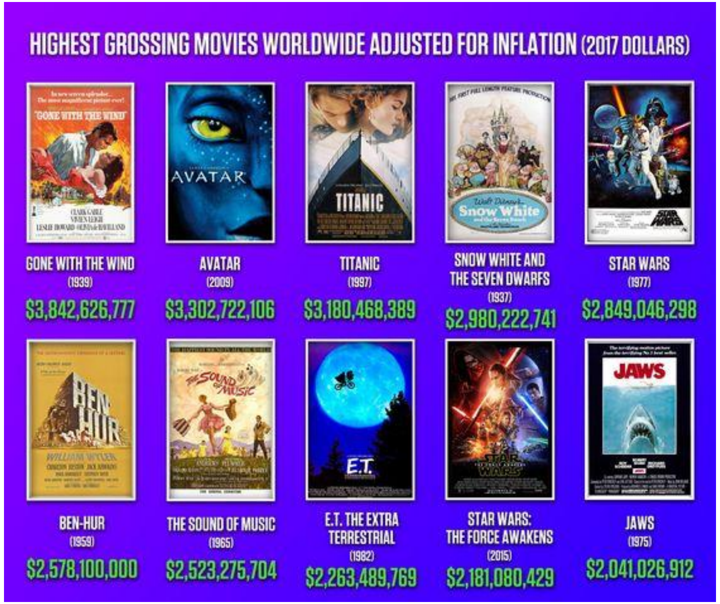
3. The Godfather: Part II
1974 3h 22min R
★ 9.0 ★ 9 90
Ways to Watch

4. The Dark Knight
2008 2h 32min PG-13
★ 9.0 ★ 10 84
Ways to Watch

5. Avengers: Endgame
2019 3h 1min PG-13
★ 8.9 ★ 10 77
View Showtimes

The factors that influence the people's decision to see a movie are genres, actors, directors, plot summary, word of mouth advertising, IMDb scores, professional movie critical reviews, IMDb user reviews, popularity, languages and production countries.

We recommend for Spectacular Studios to consider producing in the next three years, **movie projects that have the top 5 features** out of the t**op 250 movies based on IMDb score** and the **top 250 movies based on estimated profit**.



HIGHEST GROSSING MOVIES WORLDWIDE ADJUSTED FOR INFLATION (2017 DOLLARS)

GONE WITH THE WIND (1939) $3,842,626,777
AVATAR (2009) $3,302,722,106
TITANIC (1997) $3,180,468,389
SNOW WHITE AND THE SEVEN DWARFS (1937) $2,980,222,741
STAR WARS (1977) $2,849,046,298
BEN-HUR (1959) $2,578,100,000
THE SOUND OF MUSIC (1965) $2,523,275,704
E.T. THE EXTRA TERRESTRIAL (1982) $2,263,489,769
STAR WARS: THE FORCE AWAKENS (2015) $2,181,080,429
JAWS (1975) $2,041,026,912

Overview of Analysis

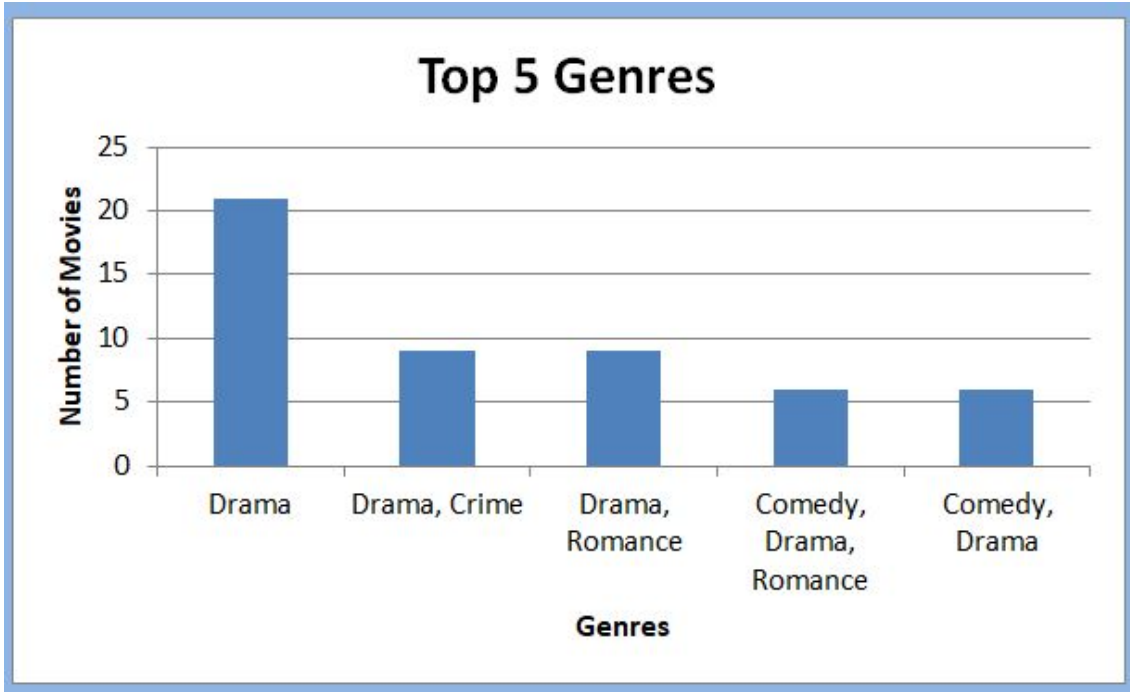
What are the key factors that influence the people's decision to see movies that belong to the **top 250 based on IMDb score** and the **top 250 based on estimated profit**?

How do these affect the recommendation to produce highly popular and financially successful movies in the next three years?

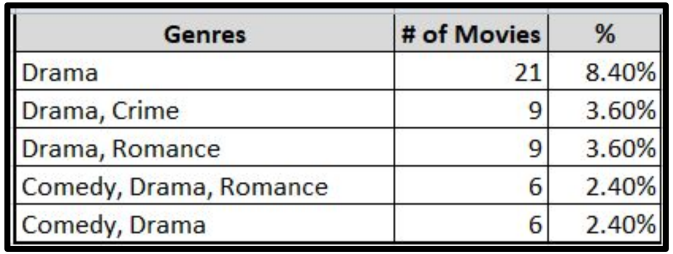
We focused our analysis on drilling down into these workstreams:

1. If **genres** influence the people to watch a movie.
2. If **popularity** influences the people to watch a movie.
3. If **original languages** influence the people to watch a movie
4. If **production countries** influence the people to watch a movie

Genre: The top 5 genres from the top 250 movies based on IMDb score
Drama on the top spot!

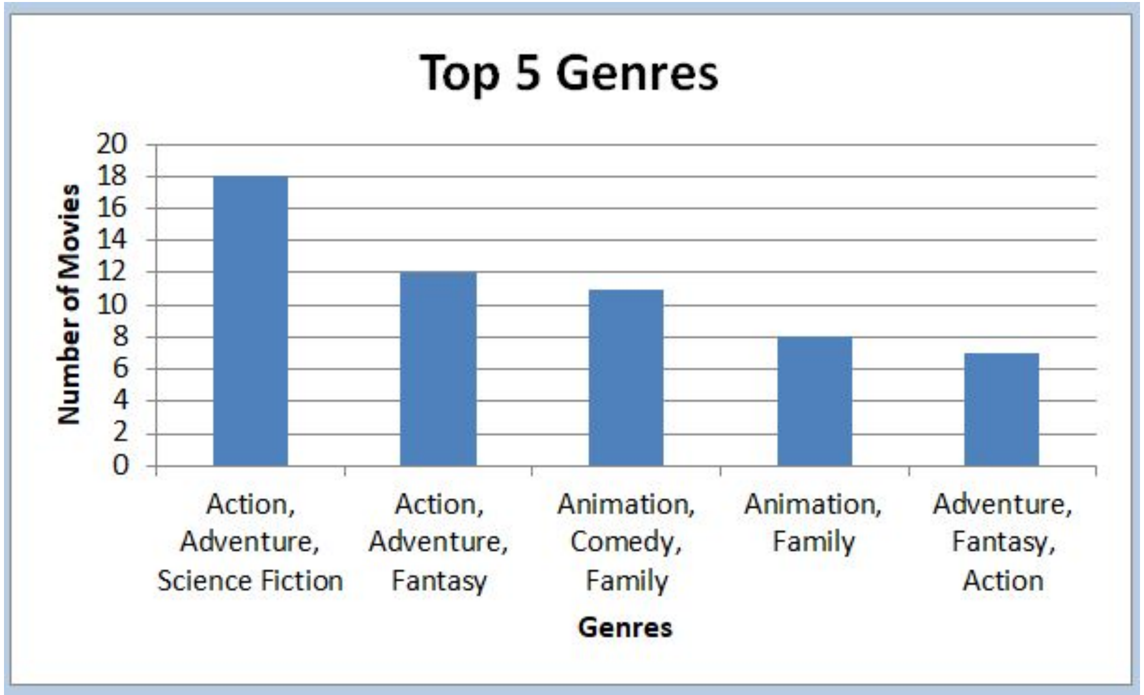


8.4% of the top 250 movies based on IMDb score are Drama, which is on the top spot. As shown in the graph and the table below, the remaining 4 are cross-genre or hybrid genre of Drama and other genres

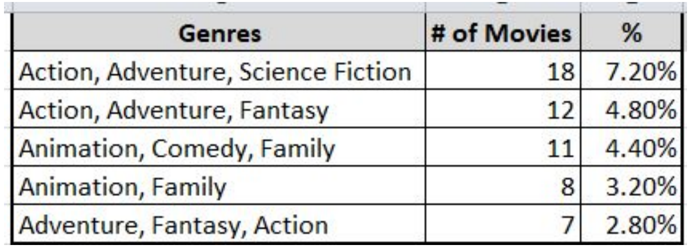


Genres	# of Movies	%
Drama	21	8.40%
Drama, Crime	9	3.60%
Drama, Romance	9	3.60%
Comedy, Drama, Romance	6	2.40%
Comedy, Drama	6	2.40%

Genre: The top 5 genres from the top 250 movies based on Estimated Profit (US$)

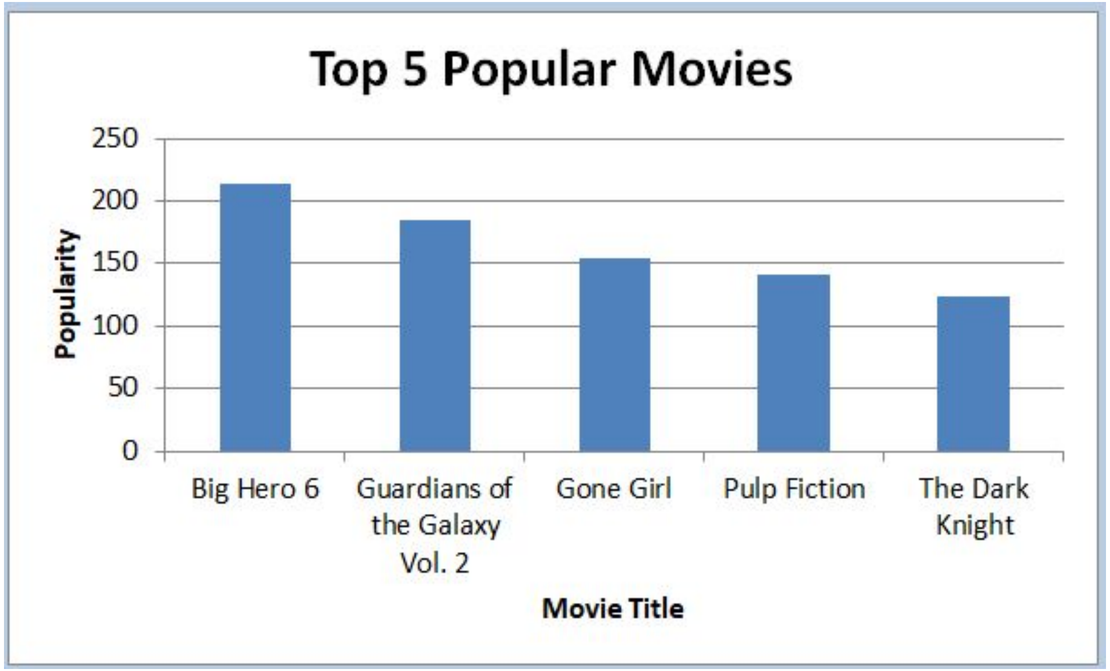


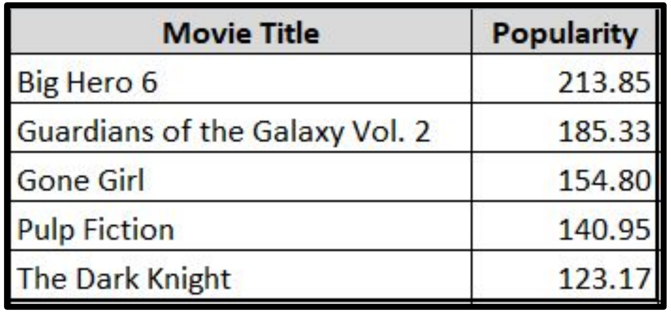
7.2 % of the top 250 movies based on estimated profit are a hybrid genre of Action, Adventure and Science Fiction. As shown in the graph and table, the remaining 4 are also cross-genres.



Genres	# of Movies	%
Action, Adventure, Science Fiction	18	7.20%
Action, Adventure, Fantasy	12	4.80%
Animation, Comedy, Family	11	4.40%
Animation, Family	8	3.20%
Adventure, Fantasy, Action	7	2.80%

Popularity: The top 5 popular movies from the top 250 movies based on IMDb score

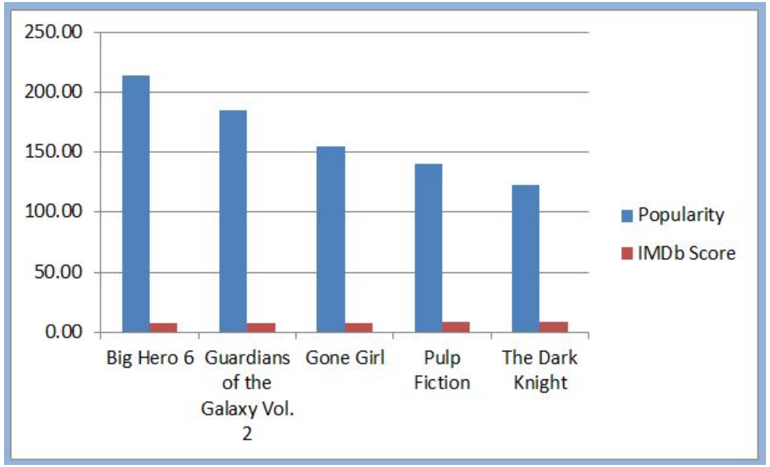


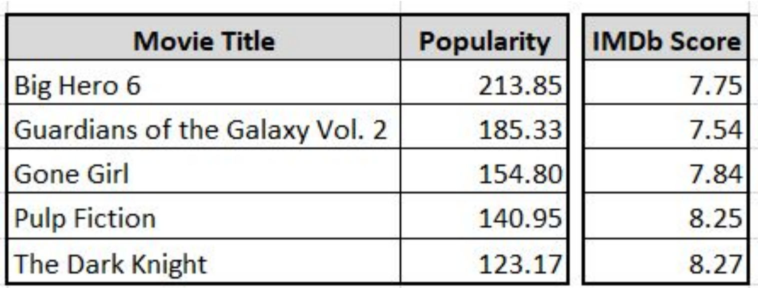


Movie Title	Popularity
Big Hero 6	213.85
Guardians of the Galaxy Vol. 2	185.33
Gone Girl	154.80
Pulp Fiction	140.95
The Dark Knight	123.17

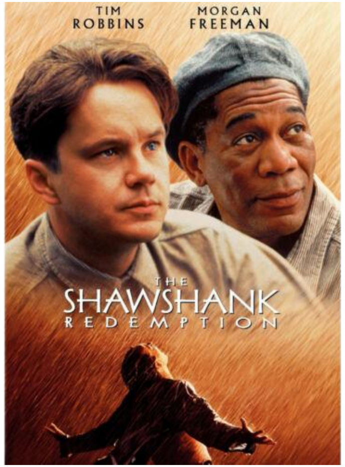
Popularity: Comparison of Top 5 movies based on Popularity and IMDb Score

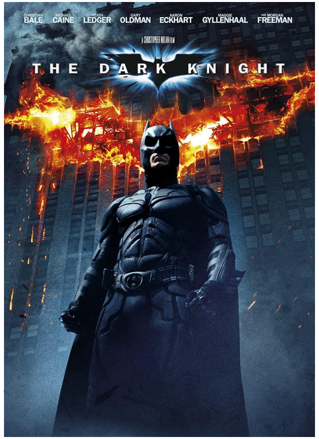
Top 5 movies based on Popularity



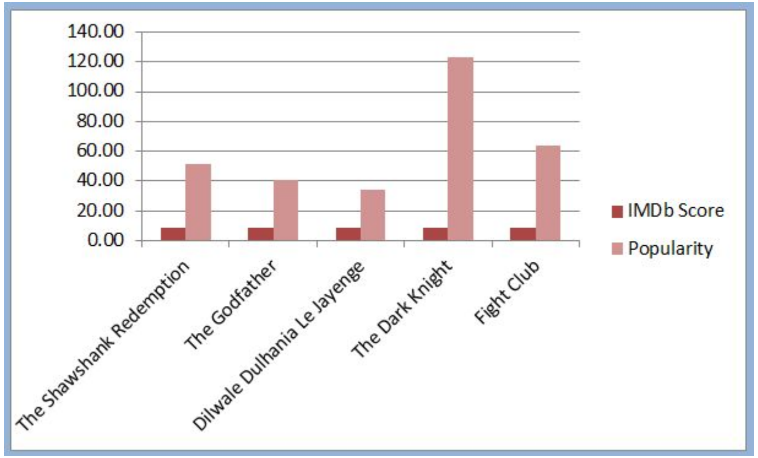


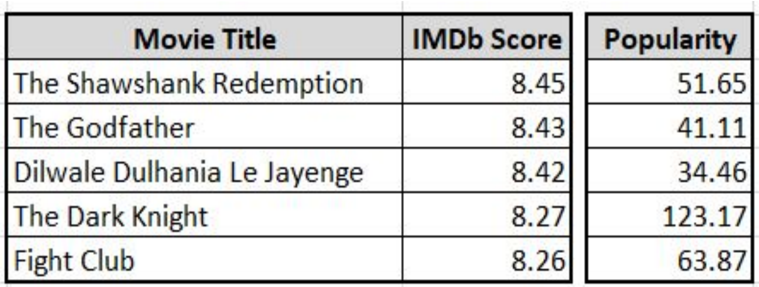
Movie Title	Popularity	IMDb Score
Big Hero 6	213.85	7.75
Guardians of the Galaxy Vol. 2	185.33	7.54
Gone Girl	154.80	7.84
Pulp Fiction	140.95	8.25
The Dark Knight	123.17	8.27





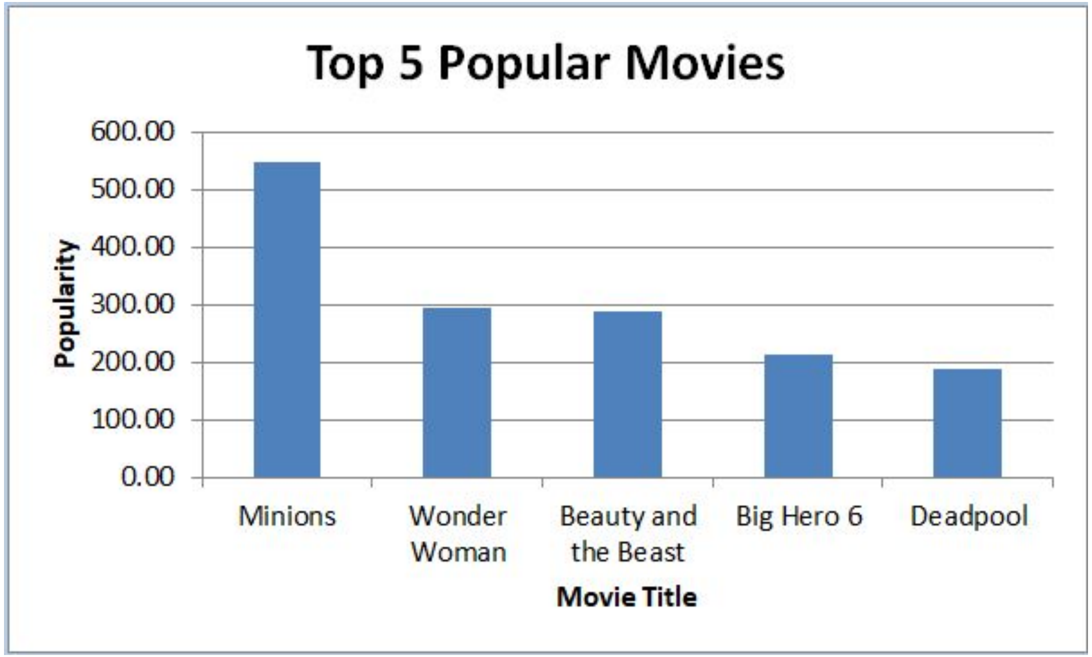
Top 5 movies based on IMDb Score





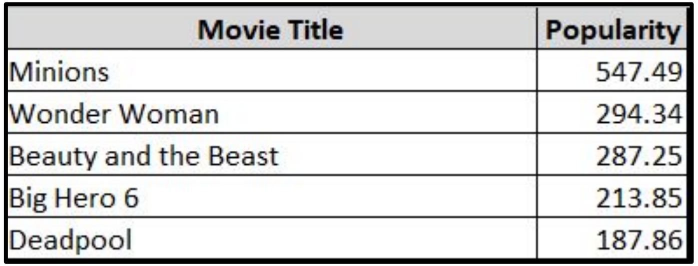
Movie Title	IMDb Score	Popularity
The Shawshank Redemption	8.45	51.65
The Godfather	8.43	41.11
Dilwale Dulhania Le Jayenge	8.42	34.46
The Dark Knight	8.27	123.17
Fight Club	8.26	63.87

Popularity: The top 5 popular movies from the top 250 movies based on Estimated Profit (US$)



ILLUMINATION ENTERTAINMENT
BOB
KEVIN
STUART
minions
SERVING EVIL SINCE THE DAWN OF TIME

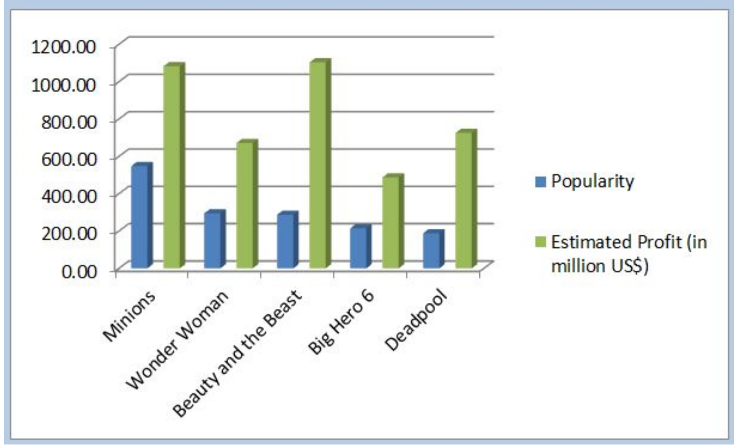
Among the top 250 movies based on estimated profit, Minions got the highest popularity score. And it shows Big Hero 6 is the 4th popular movie while it is the top popular movie based on IMDB score

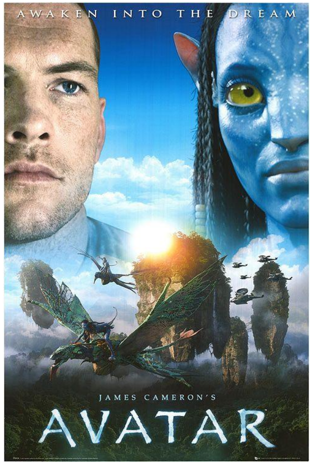


Movie Title	Popularity
Minions	547.49
Wonder Woman	294.34
Beauty and the Beast	287.25
Big Hero 6	213.85
Deadpool	187.86

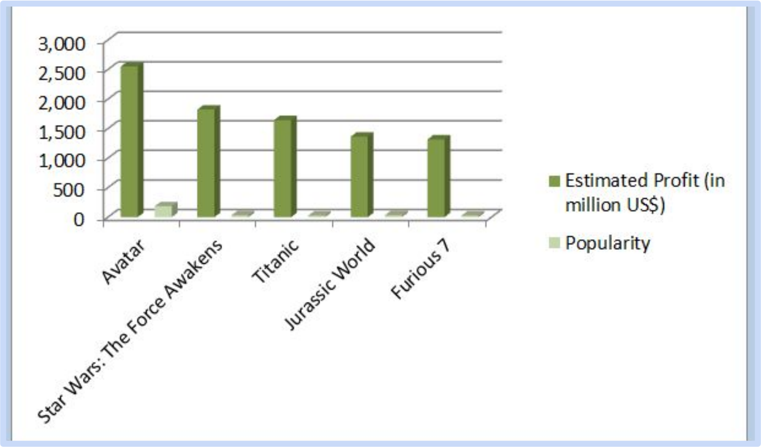
Popularity: Comparison of Top 5 movies based on Popularity and Estimated Profit (US$)

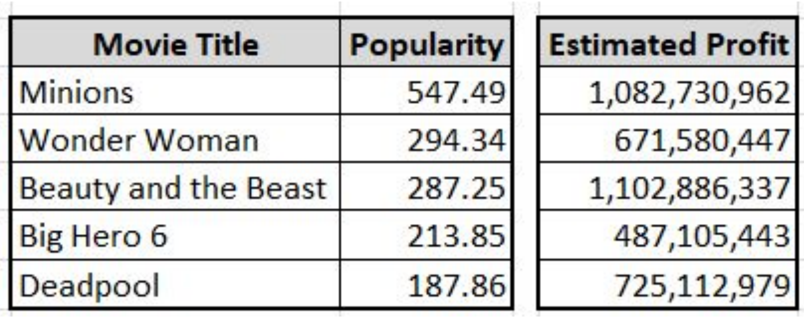
Top 5 movies based on Popularity



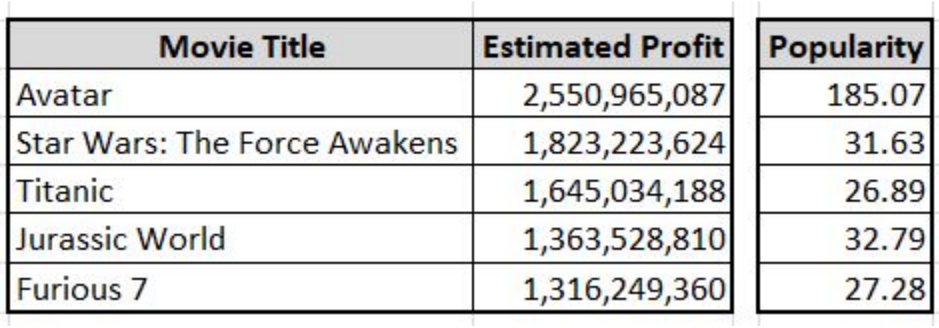


Top 5 movies based on Estimated Profit



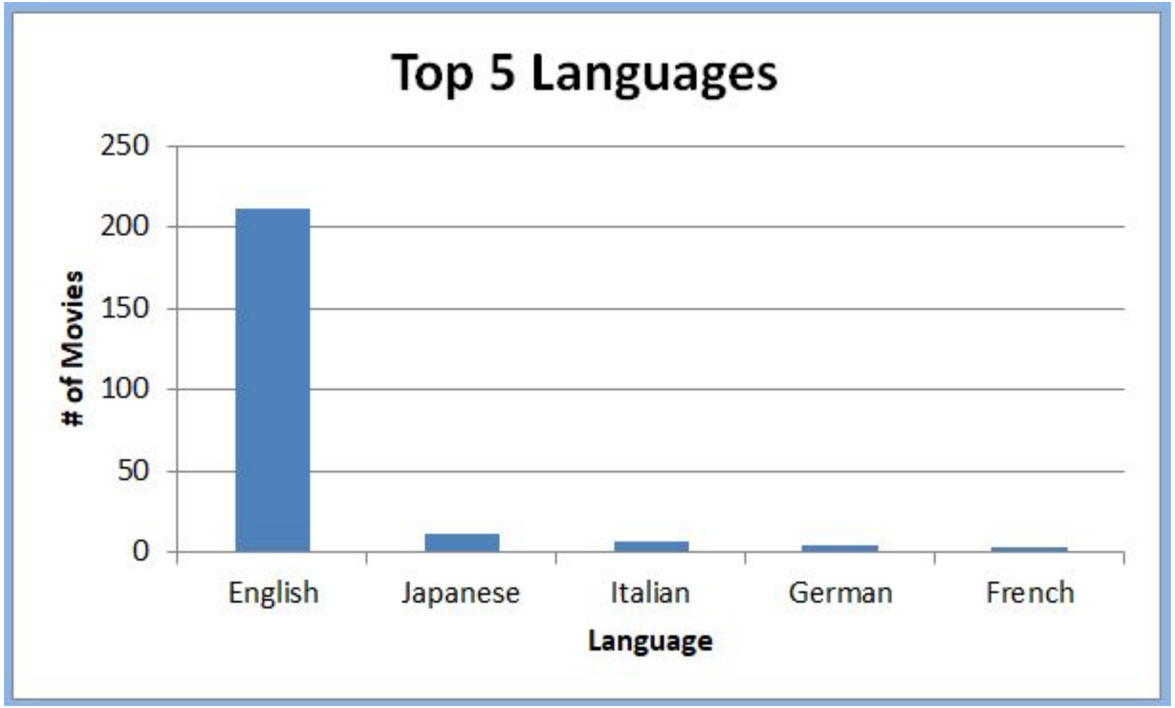


Movie Title	Popularity	Estimated Profit
Minions	547.49	1,082,730,962
Wonder Woman	294.34	671,580,447
Beauty and the Beast	287.25	1,102,886,337
Big Hero 6	213.85	487,105,443
Deadpool	187.86	725,112,979

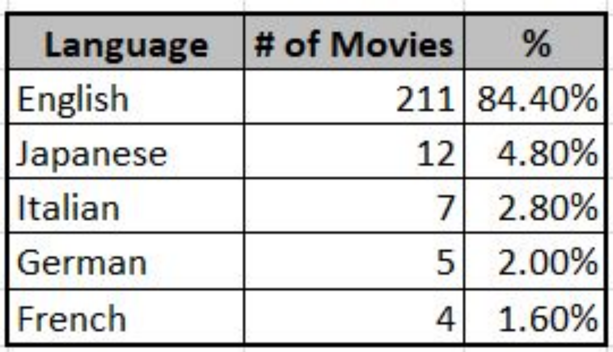


Movie Title	Estimated Profit	Popularity
Avatar	2,550,965,087	185.07
Star Wars: The Force Awakens	1,823,223,624	31.63
Titanic	1,645,034,188	26.89
Jurassic World	1,363,528,810	32.79
Furious 7	1,316,249,360	27.28

Original Language: The Top 5 Languages from the top 250 movies based on IMDB score

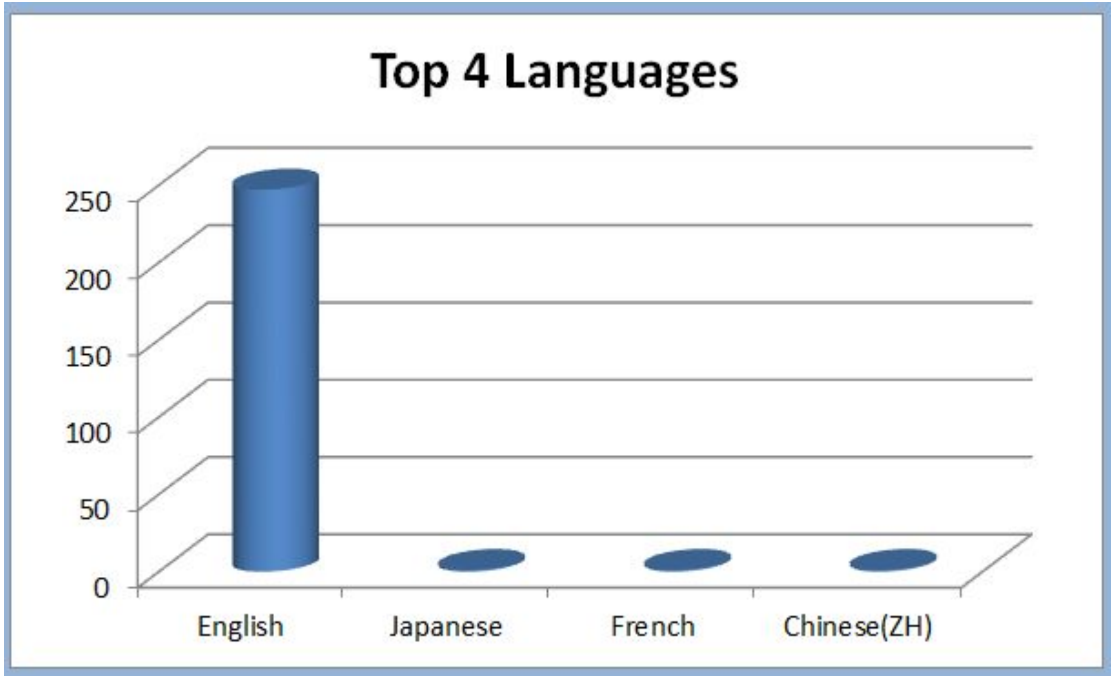


Out of the 90 languages represented in the dataset, as expected, English language movies are the most watched and most liked movies which form the 84% of the top 250 movies. Japanese and Italian came at a very distant second and third respectively.

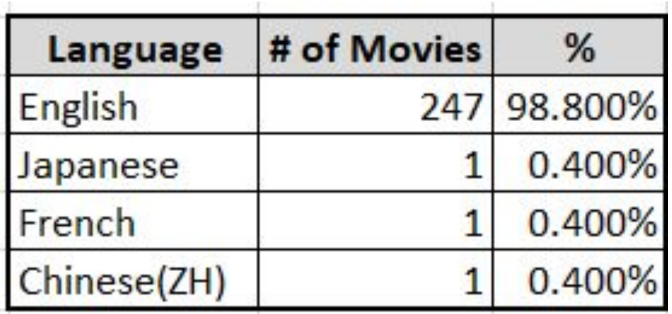


Language	# of Movies	%
English	211	84.40%
Japanese	12	4.80%
Italian	7	2.80%
German	5	2.00%
French	4	1.60%

Original Language: The Top 4 Languages from the top 250 movies based on Estimated Profit (US$)

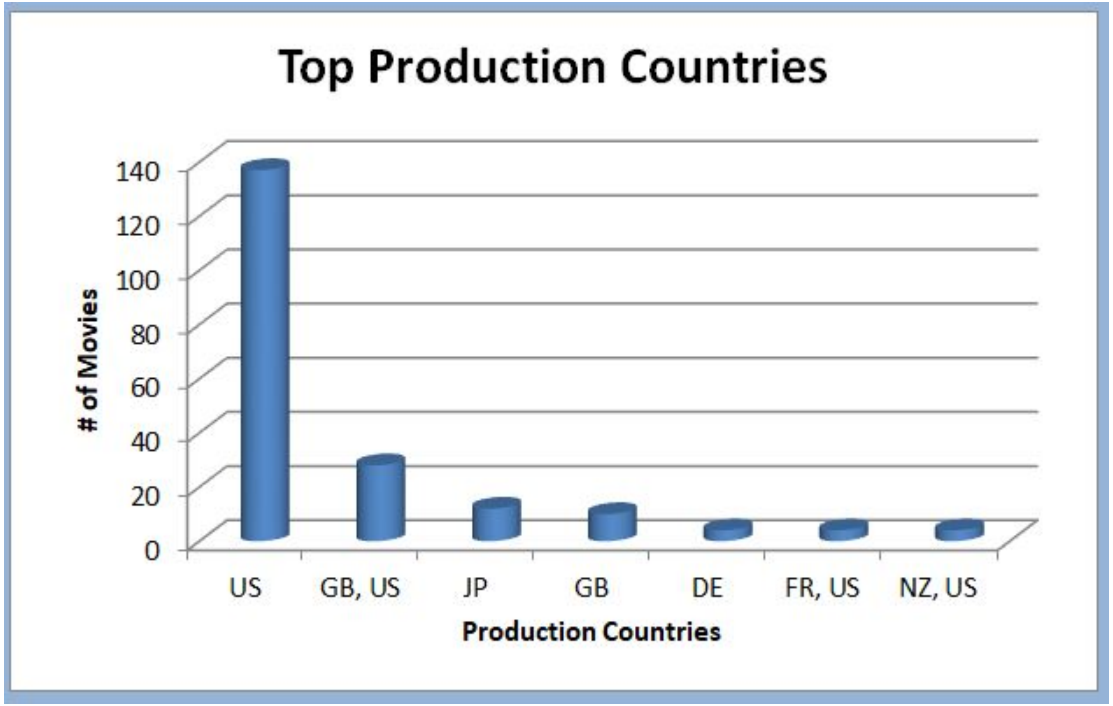


For the top 250 movies based on estimated profit, English language movies are the most financially successful movies.

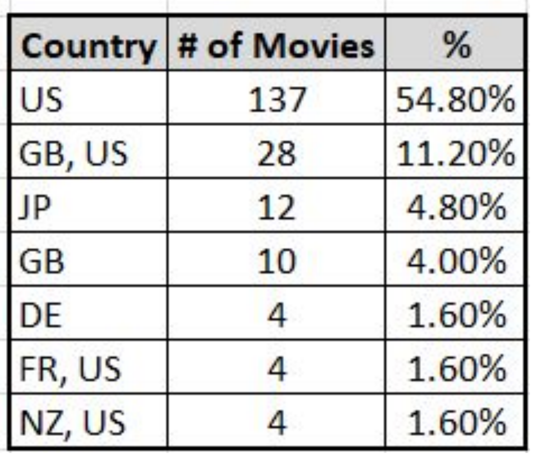


Language	# of Movies	%
English	247	98.800%
Japanese	1	0.400%
French	1	0.400%
Chinese(ZH)	1	0.400%

Production Country: The Top 5 Countries from the top 250 movies based on IMDB score

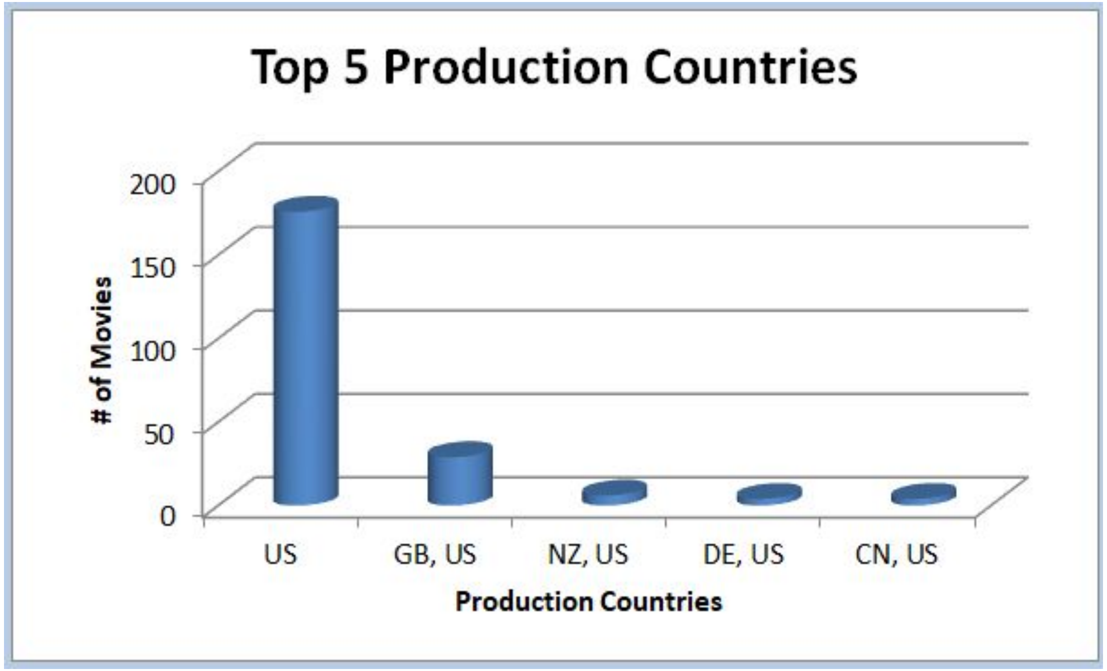


Almost 55% of the top 250 movies based on IMDb score are produced in the US. Followed by an international co-production of Great Britain and the US.

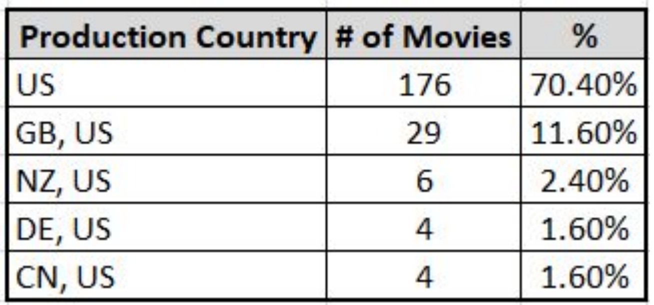


Country	# of Movies	%
US	137	54.80%
GB, US	28	11.20%
JP	12	4.80%
GB	10	4.00%
DE	4	1.60%
FR, US	4	1.60%
NZ, US	4	1.60%

Production Country: The Top 5 Countries from the top 250 movies based on Estimated Profit (US$)

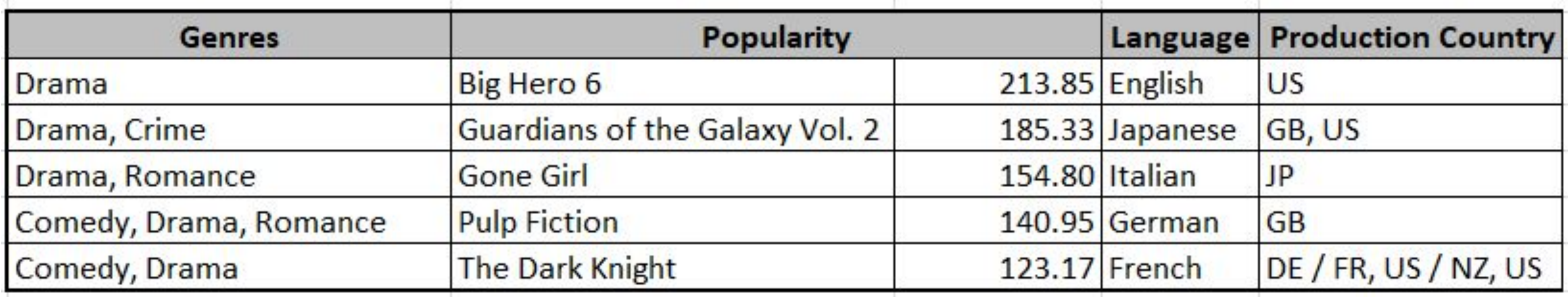


70% of the top 250 movies based on estimated profit are produced in the US. The next top 4 are international co-production between US and countries Great Britain, New Zealand, Germany and China respectively



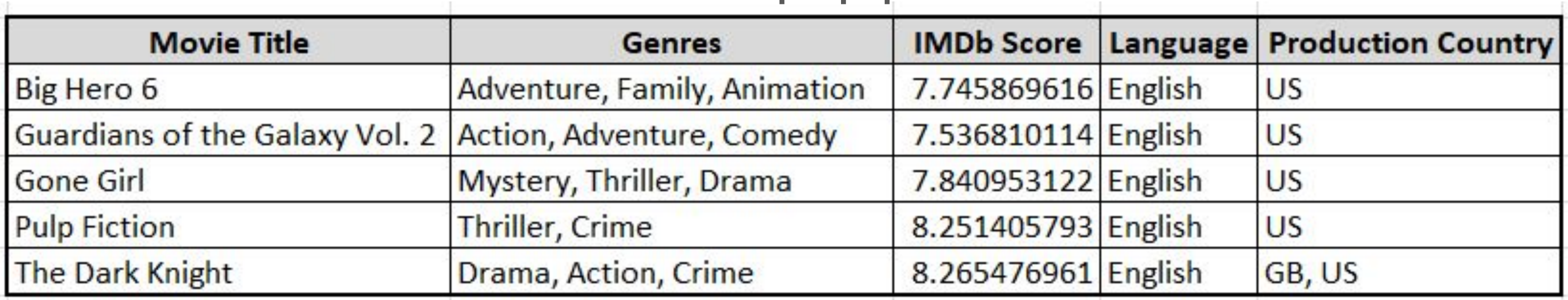
Production Country	# of Movies	%
US	176	70.40%
GB, US	29	11.60%
NZ, US	6	2.40%
DE, US	4	1.60%
CN, US	4	1.60%

Summary: Top 5 features from the top 250 movies based on IMDb score



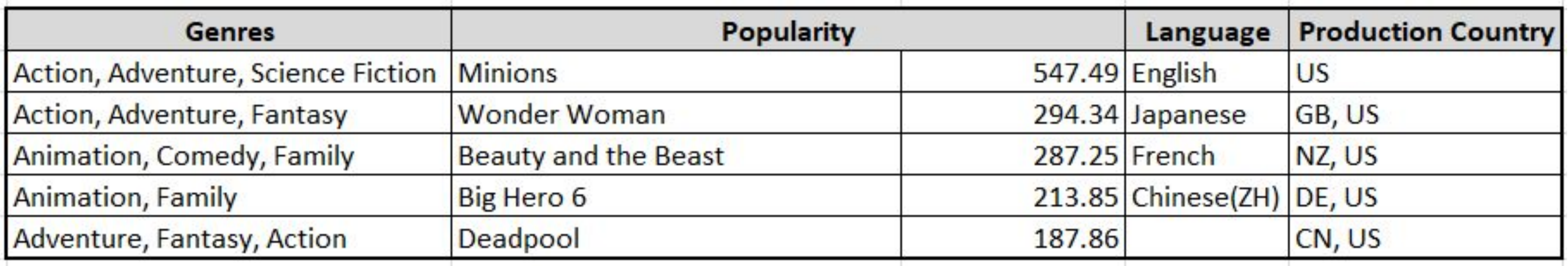
Genres	Popularity		Language	Production Country
Drama	Big Hero 6	213.85	English	US
Drama, Crime	Guardians of the Galaxy Vol. 2	185.33	Japanese	GB, US
Drama, Romance	Gone Girl	154.80	Italian	JP
Comedy, Drama, Romance	Pulp Fiction	140.95	German	GB
Comedy, Drama	The Dark Knight	123.17	French	DE / FR, US / NZ, US

The features of top 5 popular movies



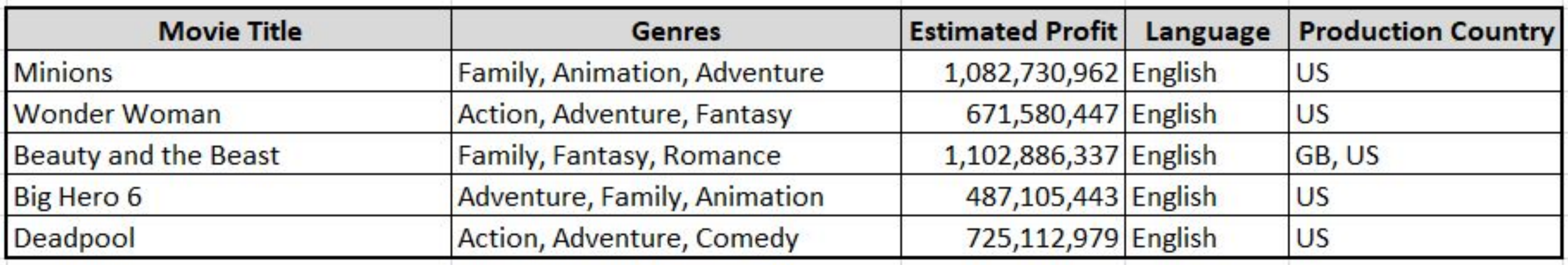
Movie Title	Genres	IMDb Score	Language	Production Country
Big Hero 6	Adventure, Family, Animation	7.745869616	English	US
Guardians of the Galaxy Vol. 2	Action, Adventure, Comedy	7.536810114	English	US
Gone Girl	Mystery, Thriller, Drama	7.840953122	English	US
Pulp Fiction	Thriller, Crime	8.251405793	English	US
The Dark Knight	Drama, Action, Crime	8.265476961	English	GB, US

Summary: Top 5 features from the top 250 movies based on Estimated Profit (US$)



Genres	Popularity		Language	Production Country
Action, Adventure, Science Fiction	Minions	547.49	English	US
Action, Adventure, Fantasy	Wonder Woman	294.34	Japanese	GB, US
Animation, Comedy, Family	Beauty and the Beast	287.25	French	NZ, US
Animation, Family	Big Hero 6	213.85	Chinese(ZH)	DE, US
Adventure, Fantasy, Action	Deadpool	187.86		CN, US

The features of top 5 popular movies



Movie Title	Genres	Estimated Profit	Language	Production Country
Minions	Family, Animation, Adventure	1,082,730,962	English	US
Wonder Woman	Action, Adventure, Fantasy	671,580,447	English	US
Beauty and the Beast	Family, Fantasy, Romance	1,102,886,337	English	GB, US
Big Hero 6	Adventure, Family, Animation	487,105,443	English	US
Deadpool	Action, Adventure, Comedy	725,112,979	English	US

Limitations and Biases

Data Collection:

1. An updated dataset is vital most importantly on the IMDb scores and Popularity scores which consistently change over time and calculations are being updated on a weekly basis.
2. There is a limitation on the fact that the IMDb's 83 million registered users are not the absolute representation of the world's total movie going audience. Not all moviegoers are IMDb registered user.
3. Other factors that affect the people's decision to see a movie are the professional movie critical review, IMDb user review, actors, directors, plot summary, word of mouth advertising which are not present in the feature set. And factors such as plot summary and word of mouth advertising are not possible to measure.

Limitations and Biases

Data Preprocessing:

1. The genres are a stringified list of dictionaries that list out all the genres and hybrid genres per movie which has about 5-6 genres. Upon data preprocessing, the genres were converted into a list of maximum of 3 genres only.

2. The production countries are a stringified list of countries where the movies are produced. Some of the movies are international co-production between 5-6 countries. Upon data preprocessing, the production countries were converted into a list of maximum of 3 countries only.

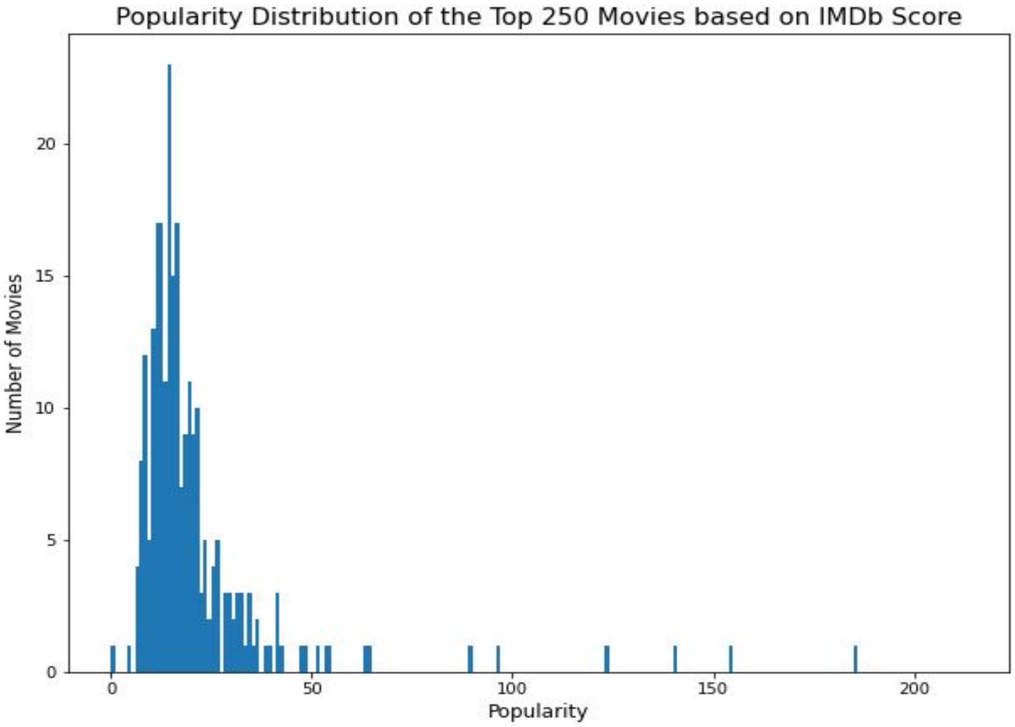
3. Less than 1% of missingness in the following features:

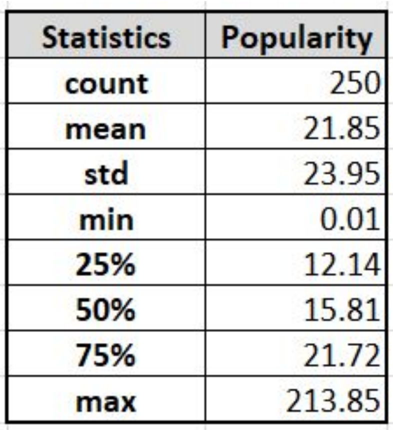
   vote_average (6), vote_count (6), revenue (6), popularity (5), language (11) and production countries (3)

   Out of the 45466 records, missing values were just around less than 1% therefore, these were just ignored due to a very small percentage.

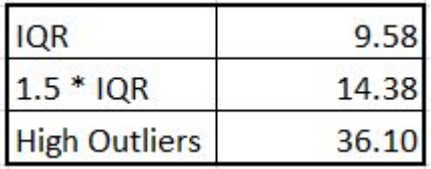
Limitations and Biases

Insights : Popularity distribution of the Top 250 movies based on IMDb Score is right skewed with 8.8% high outliers. However, these outliers have to be included as removal of the observations will have a significant effect on the analysis.





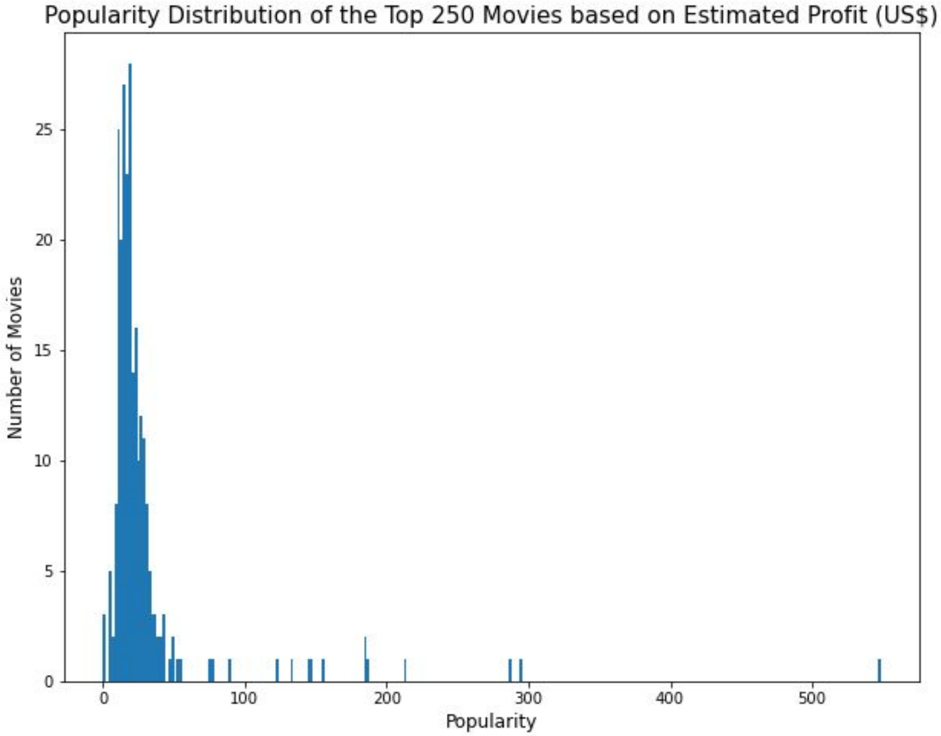
Statistics	Popularity
count	250
mean	21.85
std	23.95
min	0.01
25%	12.14
50%	15.81
75%	21.72
max	213.85

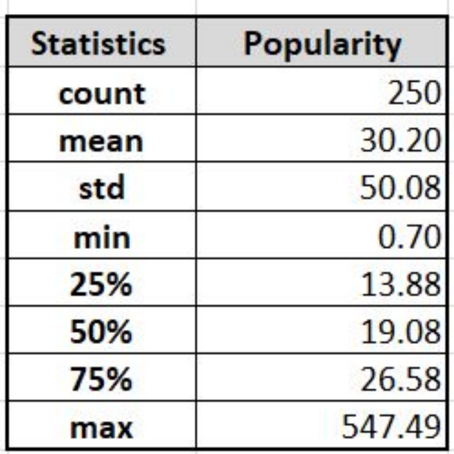


IQR	9.58
1.5 * IQR	14.38
High Outliers	36.10

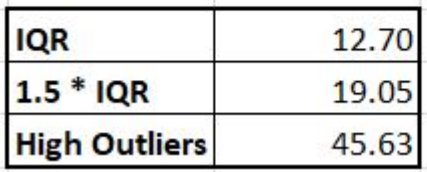
Limitations and Biases

Insights 2: Popularity distribution of the Top 250 movies based on estimated profit is right skewed with 8.0% high outliers. However, these outliers have to be included as removal of the observations will have a significant effect on the analysis.





Statistics	Popularity
count	250
mean	30.20
std	50.08
min	0.70
25%	13.88
50%	19.08
75%	26.58
max	547.49



IQR	12.70
1.5 * IQR	19.05
High Outliers	45.63

Limitations and Biases

Insights 3

IMDb uses proprietary algorithms that take into account several measures of popularity and the primary measure is what people are looking at on IMDb. IMDb records and sums the pageviews which form part of the foundation of popularity rankings.

In the feature set, the popularity is not expressed in ranking but scores which we assumed to be the number of user visits and pageviews expressed in millions.

Next Steps

1. Identify sources of potential data for popularity ranking, professional critical review, IMDb user review rating, movie actors and directors.
2. Follow-up analysis based on a more recent dataset, probably weeks-old dataset.
3. Since profitability of a film studio is crucially dependent on picking the right film projects and box office revenue is highly concentrated in a small number of very successful films, the proposed next steps from the analysis made, suggest:
    - Consider movie projects that are in the genres or hybrid genres of Drama, Crime, Romance, Comedy, Adventure, Action, Science Fiction, Fantasy and Animation
    - Produce movie projects in English language.

Thank you!