Introduction

Within the last couple of years, we have noticed that there have been a plethora of remakes. Entertainment studio Disney has been a major driver in this phenomenon, and we wanted to analyze what other movies outside of Disney have participated in this media replication and/or reimaginings. We were interested in seeing what remakes of movies have been the most popular, which ones get the most votes, and what movies have the highest average score voters.

The Dataset

Data Collection

What's in it

Columns:

- Original Language – The language the movie is originally written/performed in
- Original Title
- Overview - IMDb Description
- Popularity - A metric calculating popularity for the day (used for algorithms)[1]
- Release Date
- Vote Average - average score based off of voters
- Vote Count - # of people who voted

25,361 rows

Where it's from

Our data was split among two major sources: IMDb and TMDB.

IMDB is mostly made up of volunteers and registered users of credible standing who submit suggestions or make edits on actor, film, and television pages. The information is crowdsourced, and contributions are reviewed by IMDb employees

> "[contributors] include… actors adding their own credits; production companies filing content for their productions; and most of all, individual volunteers contributing wherever they see fit" (Lurie)[2]

[1] https://developer.themoviedb.org/docs/popularity-and-trending

[2] https://www.wired.com/story/superusers-behind-imdb-the-internets-favorite-movie-site/

> “Contributions are reviewed by IMDb, though the company is opaque when it comes to what exactly that process entails. A representative for IMDb wouldn’t share how many moderators and editors are employed by the site, nor the extent to which they may gather or revise content themselves—only that they ‘have teams and mechanisms for reviewing data to ensure it’s as accurate and reliable as possible.’” (Lurie)

IMDb recognizes notable people who have contributed to the page through their Contributor Hall of Fame[3].

TMDb works much like its more popular competitor IMDb, but on a smaller scale. TMDb is also less transparent about its moderating system, simply citing that there are employees and that the data is collected by its online community of users.

> “Where does your data come from? You! Every title since 2009 has been added by users like yourself.”[4]

How we collected it

To start the data collection process, our team investigated common keywords under popular movie remakes. By thoroughly examining these movies, we found that tags such as “remake” or “reboot” were common under these movies. In order to streamline our results, we narrowed our focus to keywords associated with over 1000 tagged movies. Once we compiled our desired keywords, we used Cinemagoer, a Python package designed for retrieving movie data from the IMDb database, to collect the respective movie IDs. With these unique IDs, we utilized the TMDB API and efficiently gathered the necessary information.

Limitations

While collecting our data, we ran into several problems with APIs and rate limitations:

1) **IMDb has an API, but only technically –** Despite their popularity as a database, IMDb’s API lacks proper documentation, making it difficult to use.
2) **Hurrah! An external package! A very slow one, at least… –** Having failed to use the IMDb API, we opted for the Python package Cinemagoer to access IMDb information. We quickly found despite bypassing IMDb’s API rate limits, Cinemagoer collected information at a speed of 5 seconds per movie. For a database including potentially hundreds of thousands of movies to collect, this meant it would take days to collect all of the data.

[3] https://contribute.imdb.com/czone/hall_of_fame
[4] https://themoviedb.org/faq/website

This means we were only able to use IMDb (through Cinemagoer) to perform the keyword search and obtain the IDs of relevant movies.

3) **Even the keywords have limits –** Some keywords included upwards of 33,000 relevant titles. Unfortunately, Cinemagoer limits the number of titles that can be collected from each key word to 10,000. This eliminated a large portion of the ids that might have been relevant to the keywords. Being sorted by popularity by default, this means that the dataset includes mainly the 10,000 most popular titles under each keyword search.
4) **TMDB just doesn't have that many movies –** With IDs collected, we turned to TMDB for information corresponding to each movie. TMDB's limitations fall on its newness and scale. Only 66% of the IDs collected from the keyword search were found in TMDB.
5) **Movies are subject to the wrong labels** - Because the information is user tagged, IMDb and TMDB users get to decide whether or not the movie is a sequel, so some movies are incorrectly labeled. The decision is subjective and not always accurate, as seen by the charts in the data analysis. Pixar's *Turning Red* is detected as a sequel, even though it's the first installment of a potential series.

In the event all of these had worked well, however there would still be some issues. Because IMDb and TMDB are essentially completely contributor based, the data included depends on individual user interest. This has led to gaps in information from non-English and non-Western media seeing as the majority of contributors are English speaking and based in the West. Someone that doesn't consume much media from these categories has a much lower chance of finding what they want on TMDB and may have to make their own contributions to the site. In general, if a piece of media is not popular enough, English speaking and western or neither, it will not be on the page – despite publisher claims of international reach.

Ethical Considerations

Ethical considerations surrounding original and remake movies research encompass a range of aspects, including preserving artistic integrity, respecting intellectual property rights, promoting representation and diversity and considering audience expectations.

The dataset contains adult content. Adult content is defined as pornographic media or adult films. Due to the nature of said films, there are major ethical and personal concerns, especially for the female identifying actors in the films. There is a history of violence towards women and people who deviate from the Anglo-Saxon cisgender heterosexual able bodied white male, and are fetishized in the content. The environments these films are produced under are unsafe and traumatizing to nearly all the actors involved, their labor is exploited, and often do not fully consent to the content they are making. By including these films in our dataset, it allows us to capture more media and content and see a more clear pattern of why remakes happen, but including these films replicates the kind of violence many people face under fetishization and the male gaze, as it allows viewers to see the adult films and give more attention to actors who may not want to remember the adult films they were a part of.

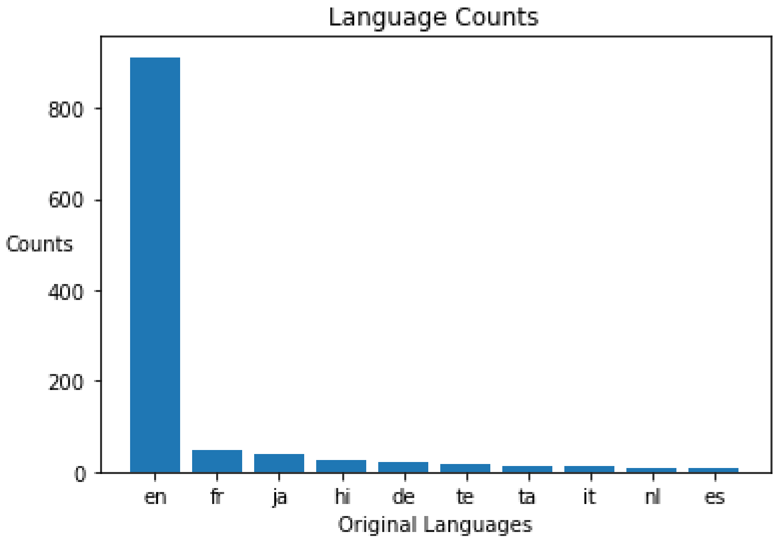
Computational Methods

a. Explain whatever methods were used for data analysis and why

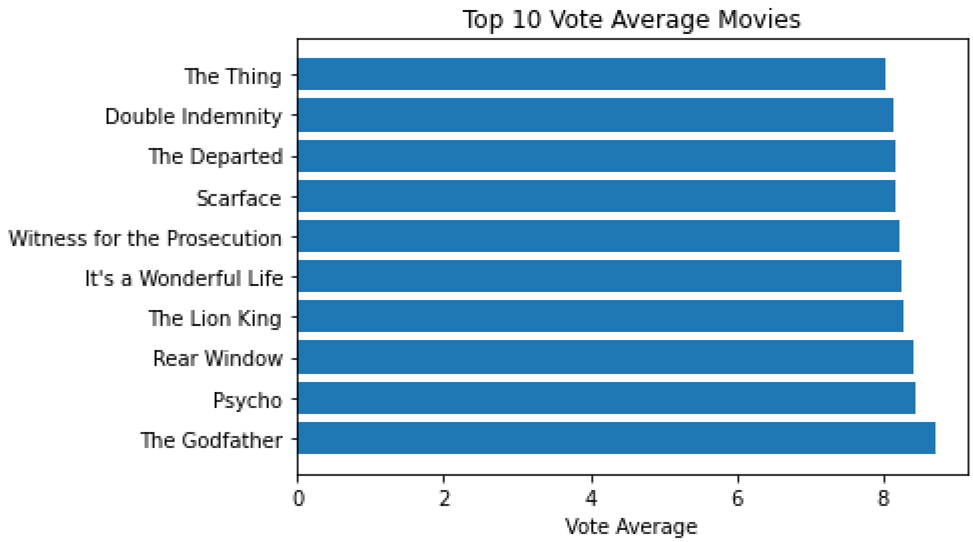
In our computational methods, we used many methods in the Python packages pandas and matplotlib to graph quantitative bar charts.

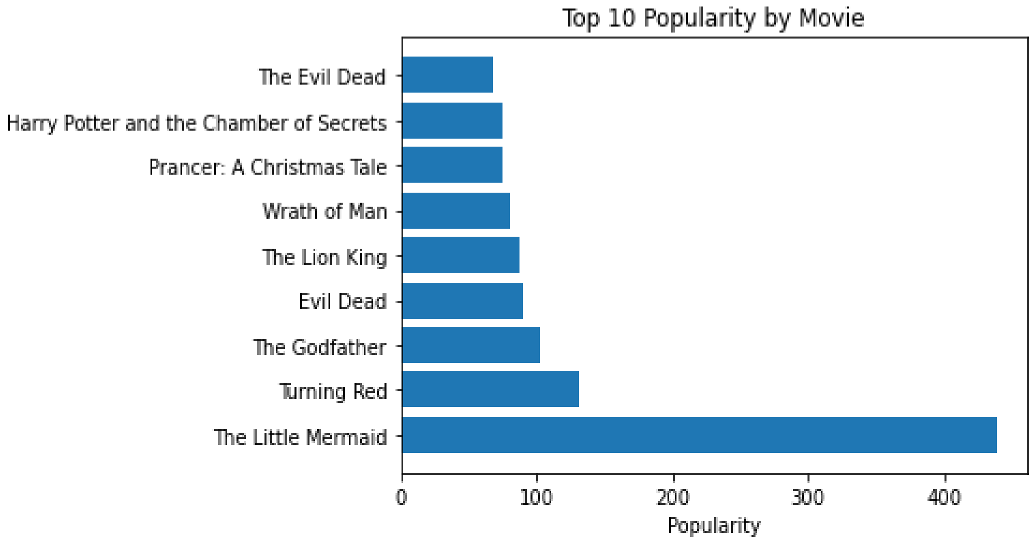
- .sample() was used to get an overview of the data and examine random rows of the data frame.
- .describe() conveniently provides descriptive statistics of the DataFrame, including both numerical and categorical columns.
- .value_counts() count the occurrences of unique values in a column of a data frame.
- .head() Show the first n rows of the dataframe.
- .sort_values() Sort the data frame by one or more columns. This helps with quickly finding information under the same column if we wanted to isolate certain categories and give a magnified observation.
- .iterrows() iterate through each row to find the matching columns.
- .loc() access and modify a data frame based on label-based indexing. The label based indexing helps us sift through data more easily.
- plt.bar() create a bar plot and plt.barh() create a horizontal bar plot. Both are used to provide a visual representation of our findings.

Data Analysis



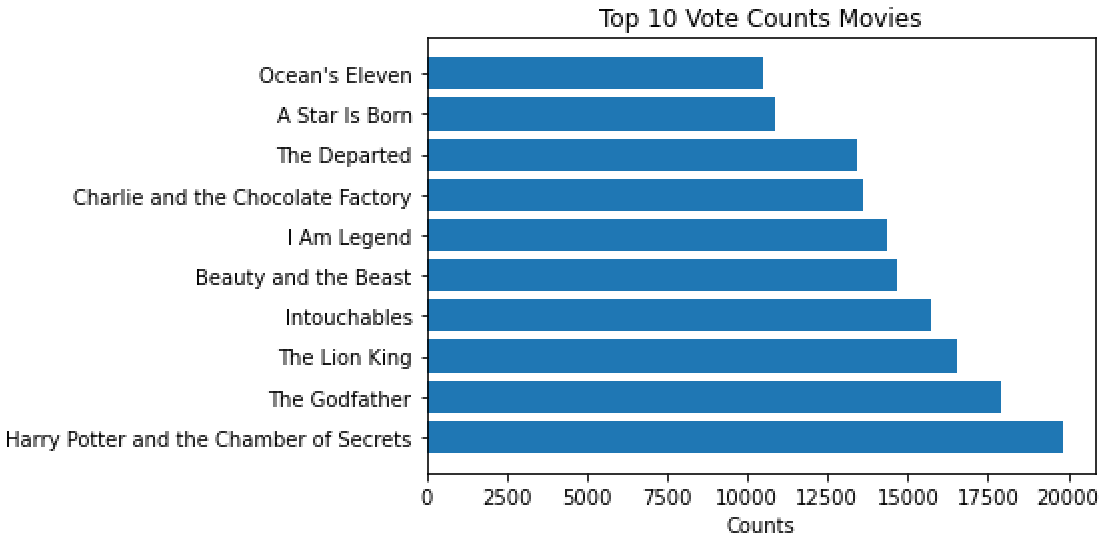
Some main findings we had were that the majority of the films included in the data set are originally spoken in English. The count reveals that most of the entries of interest are English films, and by extension, most users are English speaking, which is what we expected before officially crawling our data.

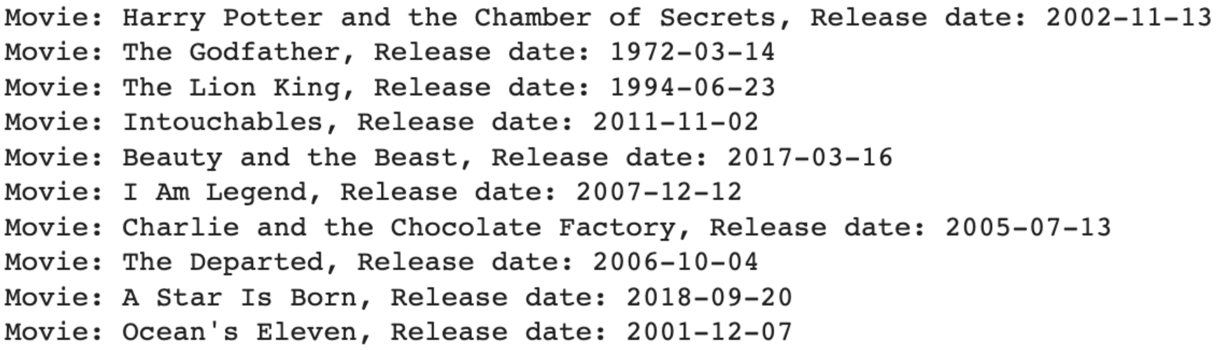




The most popular or well received movies were a mixture of genres and intended audience. The most popular are children's movies, adult dramas, and horror. It's representative of real-life patterns of what is considered safe and reliable by investors [5], especially when it comes to children's movies as the parents most likely have a positive association with the movie and want to connect with their child through watching the remake with them.

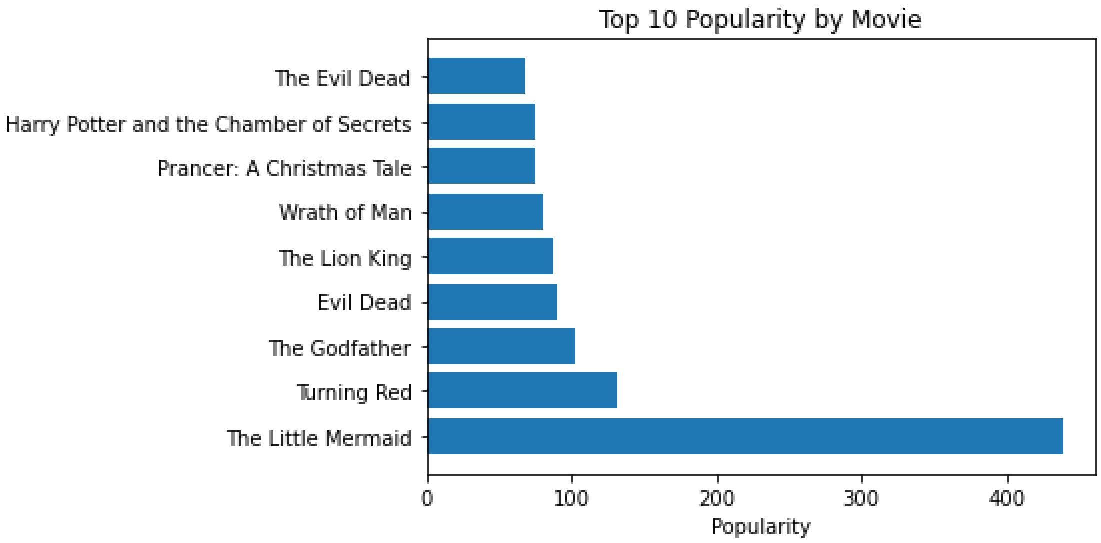
[5] https://www.vulture.com/2014/02/4-reasons-why-hollywood-still-makes-remakes.html

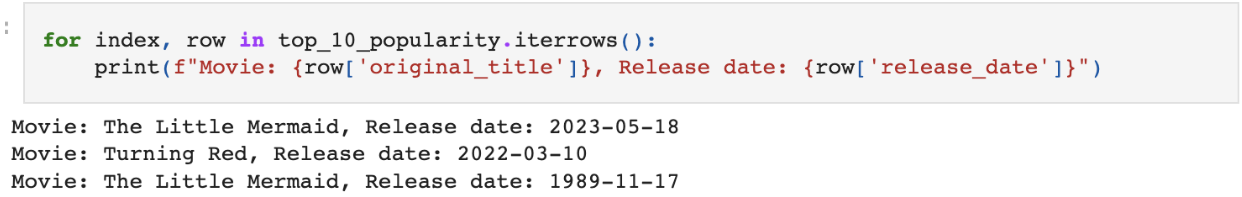




```
Movie: Harry Potter and the Chamber of Secrets, Release date: 2002-11-13
Movie: The Godfather, Release date: 1972-03-14
Movie: The Lion King, Release date: 1994-06-23
Movie: Intouchables, Release date: 2011-11-02
Movie: Beauty and the Beast, Release date: 2017-03-16
Movie: I Am Legend, Release date: 2007-12-12
Movie: Charlie and the Chocolate Factory, Release date: 2005-07-13
Movie: The Departed, Release date: 2006-10-04
Movie: A Star Is Born, Release date: 2018-09-20
Movie: Ocean's Eleven, Release date: 2001-12-07
```

A specific example that we want to bring up is the movie *A Star Is Born* showing up in the chart “Top 10 Vote Counts Movies.” The movie has had multiple remakes, but the one that shows up here is the 2018 that pop singer Lady Gaga starred in with her co-star Bradley Cooper. Past iterations of *A Star Is Born* include starring actresses Judy Garland and Barbra Streisand in their respective versions. Despite their star power, it doesn’t appear that these past versions showed up (other data visualizations will show more than one version of the movie if prompted as seen by the *Little Mermaid* showing up twice in “Top 10 Popularity by Movie”).



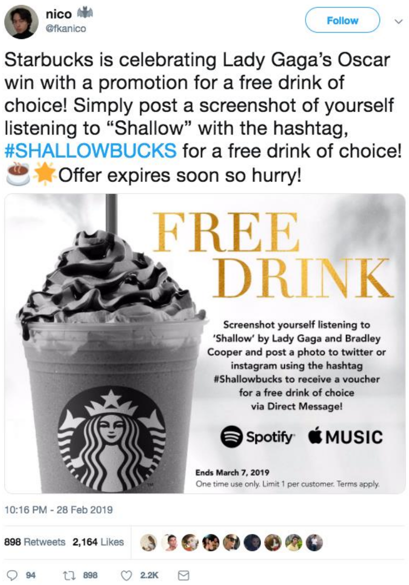


```
for index, row in top_10_popularity.iterrows():
    print(f"Movie: {row['original_title']}, Release date: {row['release_date']}")
```

```
Movie: The Little Mermaid, Release date: 2023-05-18
Movie: Turning Red, Release date: 2022-03-10
Movie: The Little Mermaid, Release date: 1989-11-17
```

This is most likely the case due to TMDB's smaller database, but also signifies the kind of modern cultural relevance Lady Gaga has in entertainment and the masses of people who consume her media. More Internet literate people, mostly made up of young adults, are going to vote on *A Star Is Born* than older fans of Judy Garland and Barbra Streisand who aren't on the Internet as much. Moreover, it shows the phenomenon of fan contributions related to this film, and fan's desires to propel Lady Gaga to even more recognition than what she already has. One of the singles off of the *A Star Is Born* soundtrack, *Shallow*, was widely circulated in a Twitter disinformation campaign called #SHALLOWBUCKS. Twitter user @fkanico, created a poster detailing that if someone streams *Shallow*, they can redeem a free Starbucks drink.[6]

[6] The Starbucks Scam that Changed Twitter Forever



nico

@fkanico

Starbucks is celebrating Lady Gaga’s Oscar win with a promotion for a free drink of choice! Simply post a screenshot of yourself listening to “Shallow” with the hashtag, #SHALLOWBUCKS for a free drink of choice! ☕🌟Offer expires soon so hurry!

FREE DRINK

Screenshot yourself listening to 'Shallow' by Lady Gaga and Bradley Cooper and post a photo to twitter or instagram using the hashtag #Shallowbucks to receive a voucher for a free drink of choice via Direct Message!

Spotify MUSIC

Ends March 7, 2019

One time use only. Limit 1 per customer. Terms apply.

10:16 PM - 28 Feb 2019

898 Retweets 2,164 Likes

Multiple fans retweeted the image to get unknowing listeners to stream the song in order to boost the popularity of the single and to keep it on the music charts. It’s a reflection of intense fan culture in the 21st century.

- Main findings
- Insights, summary stats, EDA
- Data visualization
- What viz was used, encodings, explanation of how it was made and why
- Specific Example

Future Direction

Although we came into this project with many aspirations and feelings of excitement, we were cut short by our limitations and time restraint. In the future, we are hoping to find another source to crawl our information more efficiently and quickly to be able to capture everything that we want.

We would like to clean the data and take out adult content as we are not particularly interested in that niche. Besides cleaning the data, we want to have a way of tagging what is the original source material, and definitively expand to books and plays.

We want to be able to crawl for the movie studio for each film and see growth within each studio and link them to parent companies to see more complex patterns. If we could attach box office earnings to each movie entry, then we can quickly compare it to the source material and the remake. Furthermore, if we could capture the genre of each film, we could see a pattern in popularity over time for the given genre to observe any changes or consistencies.

There are plenty of directions we can go in, and it is something we want to explore even after this class is over.