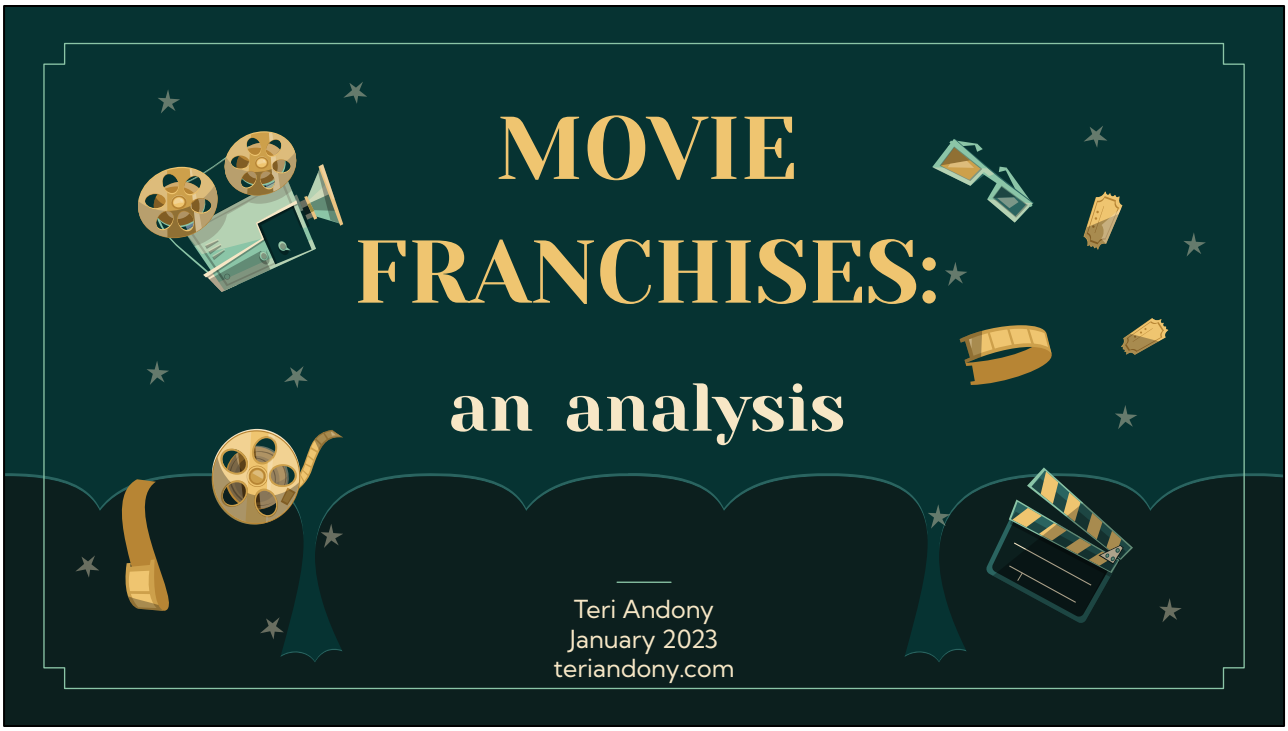
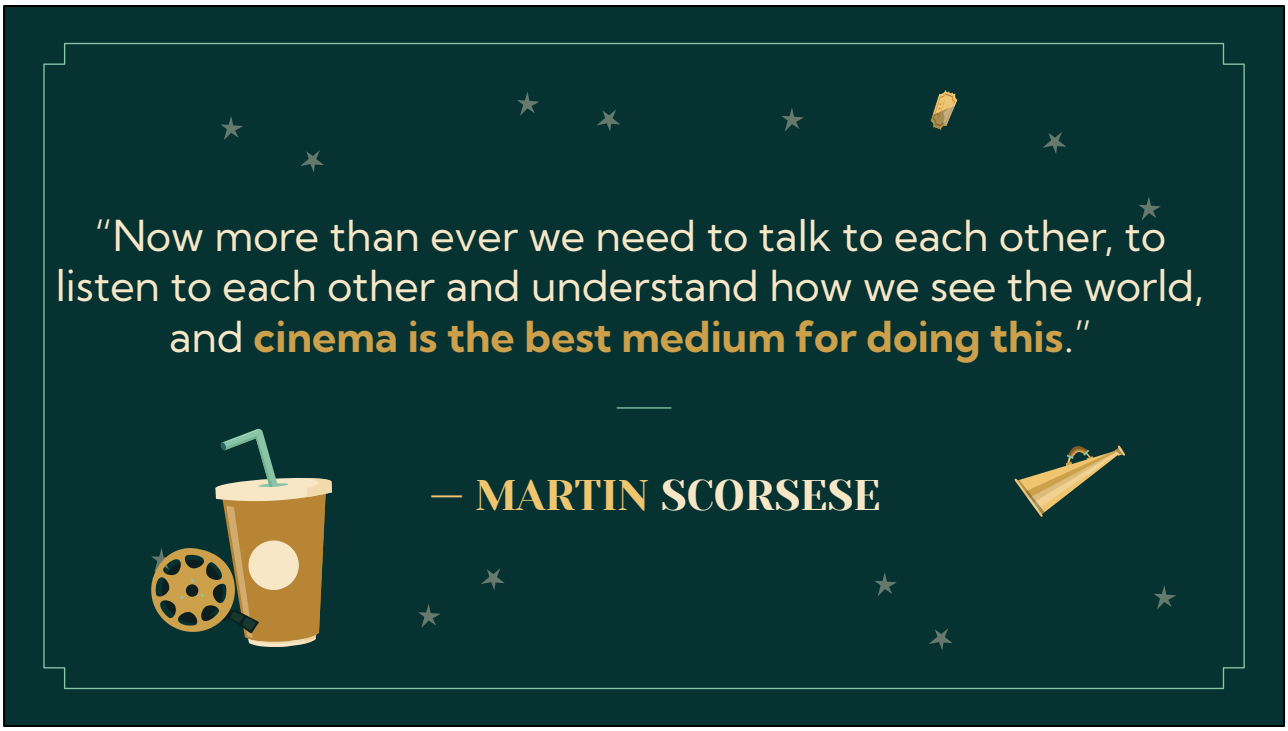

MOVIE FRANCHISES:
an analysis
Teri Andony
January 2023
teriandony.com

I'll kick things off with Part One: Origins. This is a brief introduction to my topic as well as a bit of context. Then, in Part Two: The Harvest, I'll summarize the steps I took to extract, transform, load, and analyze the data. In part 3, The Reckoning, I'll go over a few key insights I found. Part Four, The Beyond, wraps up this series with recommendations for future analysis. Aaaaaand, action!


I.
ORIGINS
Background and Research Question



You are all most likely aware of what movies are. As much as I would love to blab for hours, I'm not going to cover the history of cinema. What I would like to point out, however, is *why* I believe movies are such a vital element of society and an important topic to research. As Martin Scorsese pointed out, cinema is the best medium for connecting with each other and understanding diverse viewpoints. It's a wide-reaching medium that represents an increasingly wide spectrum of thoughts and opinions. Developing technology allows movie-making to reach more and more of the global population, and I believe it will only continue to flourish as we begin to hear from every corner of the world. Ok, soapbox over.

Franchise, *defined*

=

I chose to focus on franchises for my research. Put simply, franchises (or collections), are stories that span multiple movies, often including the same characters and similar storylines. They've been around nearly as long as movies themselves, but rose to popularity in the early 70s with the release of James Bond, Star Wars, and Planet of the Apes. They represented an important shift in mindset for the industry. The term itself, a "franchise", implies a series of products (or films) that could generate profit. As much as I would like to maintain that movies are purely an art form, it is important to recognize that this industry is also a business. It is therefore critical to understand the factors that play into what makes up a successful franchise, so that we may optimize their utility in the future.

Success, *defined*

Revenue

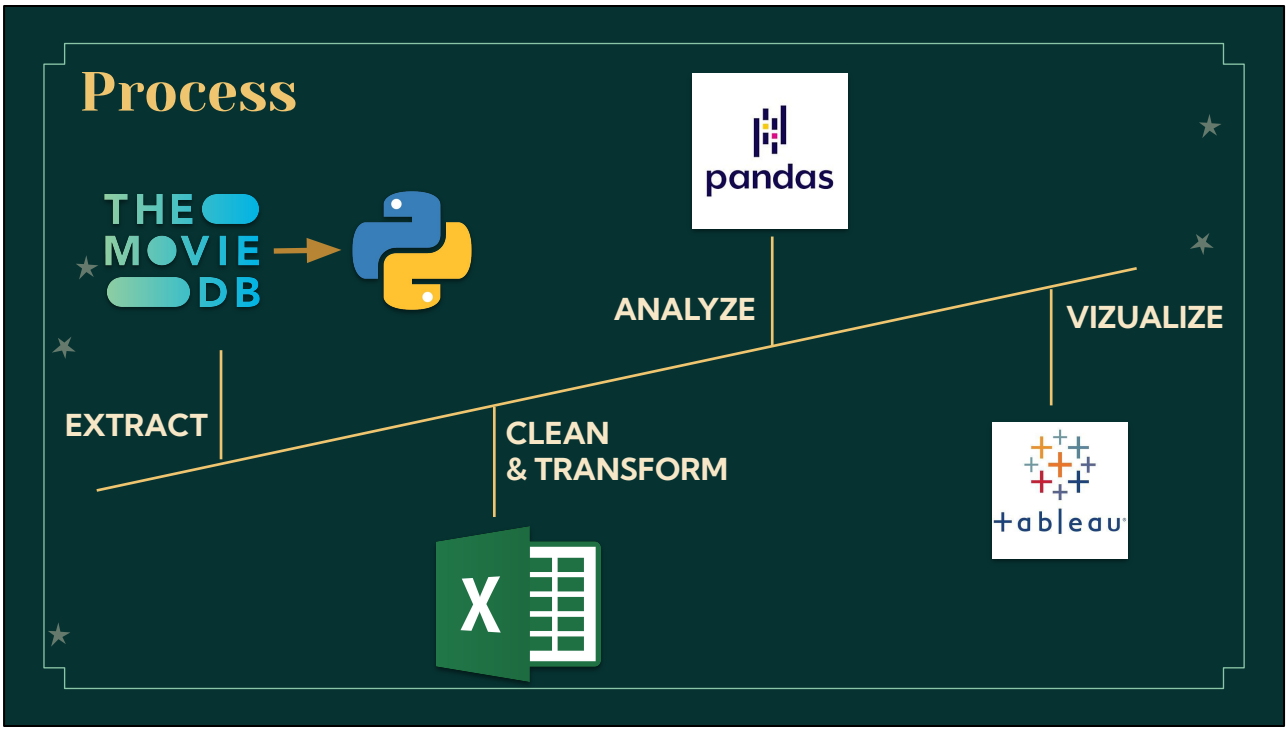
= Vote

When I set out to understand what makes up a successful franchise, I wanted to establish a clear definition of success. I considered a variety of variables, such as vote count, popularity, and rating. I ultimately settled on revenue, or gross, as my primary success metric. For our purposes, “revenue” is defined as the cumulative total of worldwide box office ticket sales generated for a film’s entire run. Though reported by the user, TMDB recommends using sources like Box Office Mojo.

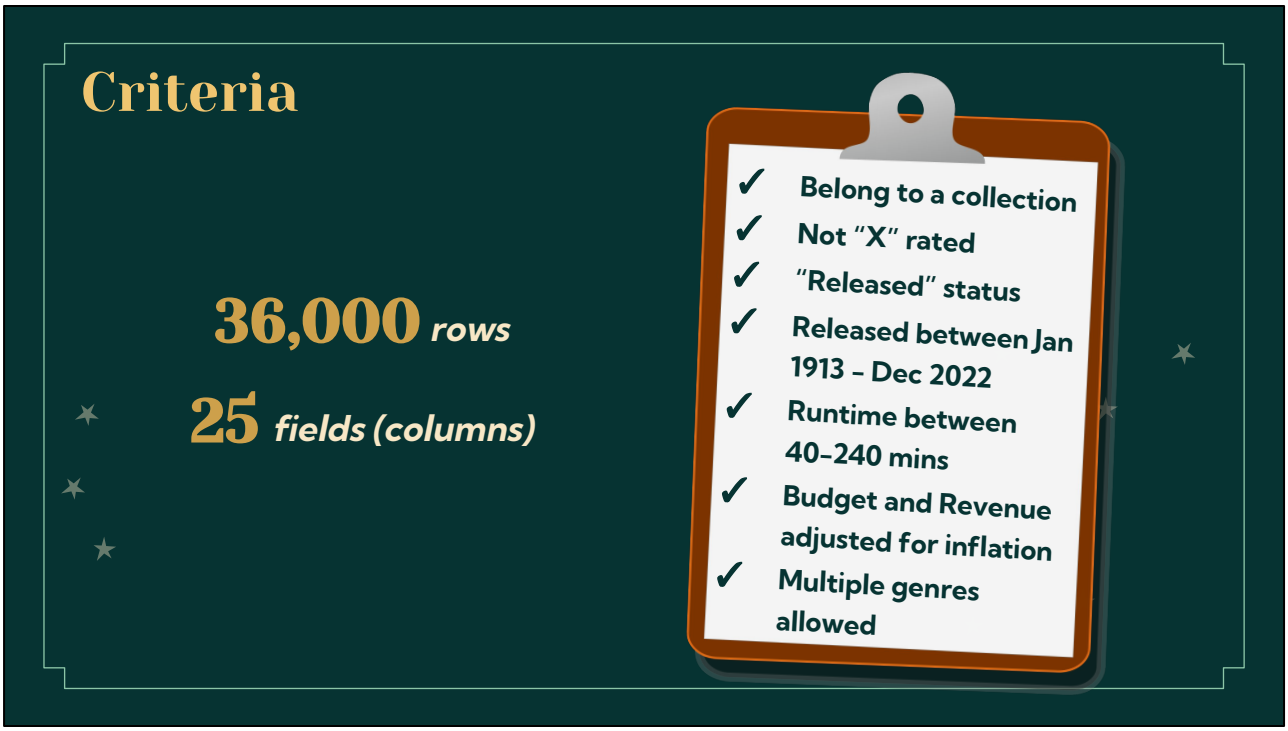
For obvious reasons, how much a movie makes is a clear indicator that it is successful. If a movie makes a lot of money, it is understandable that the creators will want to replicate their success with sequels. On a higher level, a movie’s gross represents its popularity. Purchasing a ticket to see a film is ultimately a vote for the collective body that made the film. This vote, more than the votes represented on sites like Rotten Tomatoes or The Movie Database, represents the *entire population of consumers*.

Now that we have set the stage, let’s explore the data.


II.
THE HARVEST
Data summary and exploration



I chose to extract data using an API from The Movie Database, a Canadian-founded, user-led review site. I used Python to extract the data (a process that was complicated by the fact that the API only allowed calling the information for one movie at a time). Once I had the data in hand, I imported it into Excel to clean and transform. I then imported the cleaned dataset back into Python for analysis using Pandas (I also utilized pivot charts in Excel). Finally, I created visualizations in Tableau.

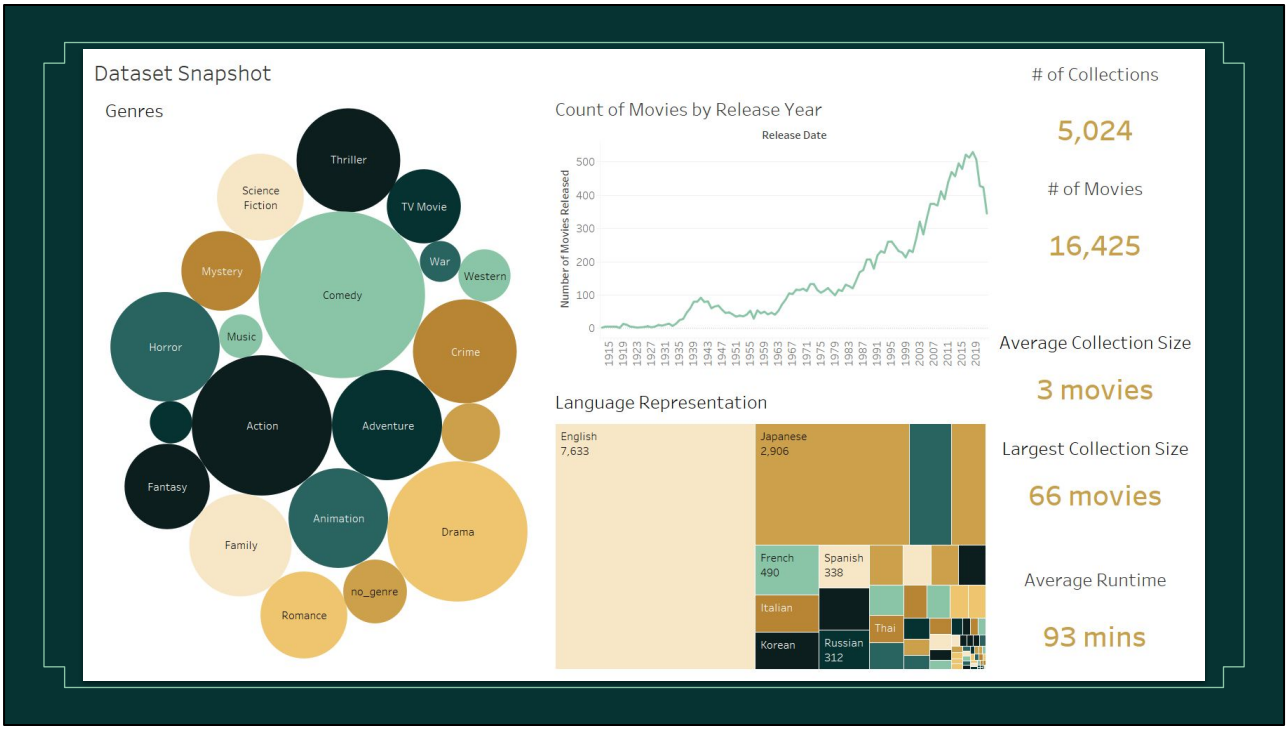


In order to qualify for inclusion in my dataset, a film had to meet the following criteria: Most importantly, it had to belong to a collection. All “adult” films were removed. The films had to have “released” status (rather than in production, cancelled, etc). The release date fell between January 1913 and December 2022. All films had a runtime of between 40 and 240 minutes to exclude short films and volumes.

I transformed the “budget” and “revenue” fields in order to adjust for inflation. I used Consumer Price Index data from Nasdaq Data Link to calculate my adjustments.

Finally, it is important to note that movies could list multiple genres. For instance, Star Wars reported 3 genres: action, adventure, and sci-fi.

I ultimately ended up with approximately 36,000 rows and 25 fields (columns).



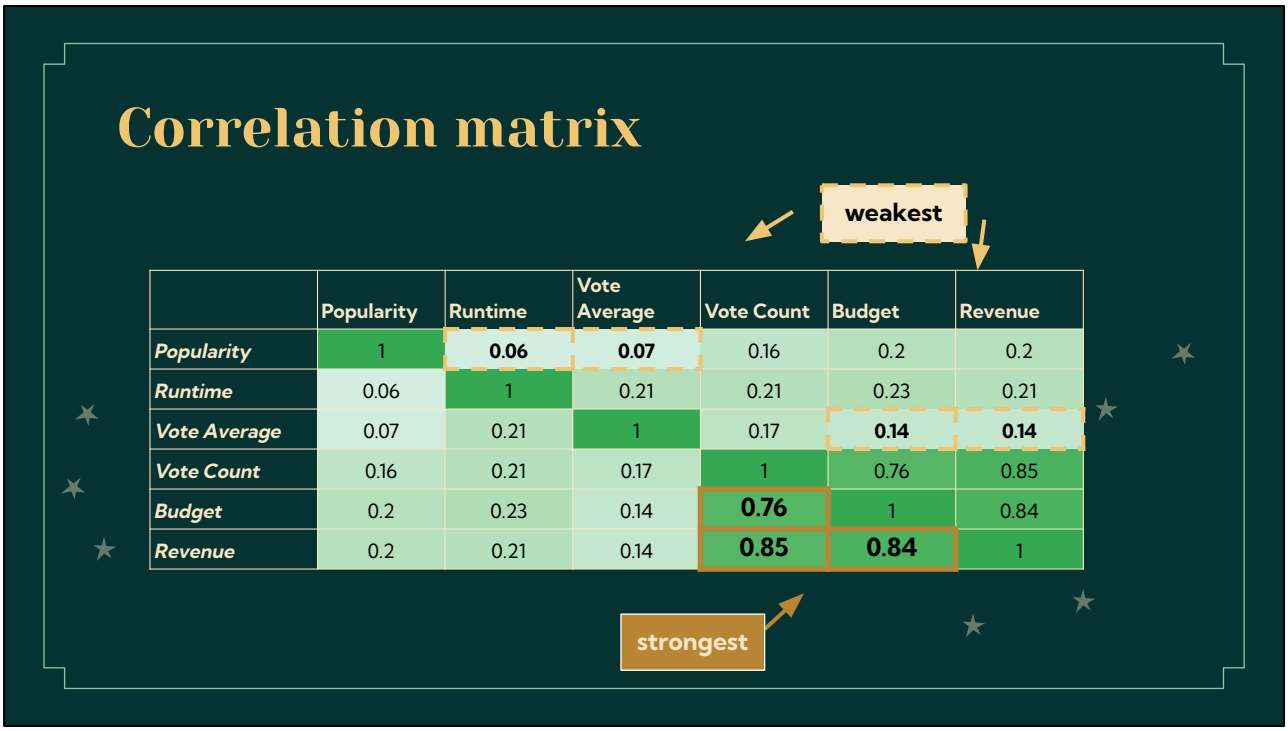
Here is a brief snapshot of the dataset.

- There were 5,024 collections comprised of 16,425 movies.
- The average collection size was 3 movies.
- The largest collection included 66 movies.
- The average runtime of these movies was 93 minutes.
- As you can see, the count of franchise movies has greatly increased, particularly since the turn of the century.
- The movies ranged across 20 genres. Comedy had the most movies, followed by Drama and then Action. War was the least represented genre.
- Finally, the movies spanned 66 languages. English was the most represented with just under 7700 films, followed by Japanese with around 3000.

I'll dive deeper into budget, revenue, popularity and rating in the next section.


III.
THE RECKONING
Analysis and results

Correlation matrix

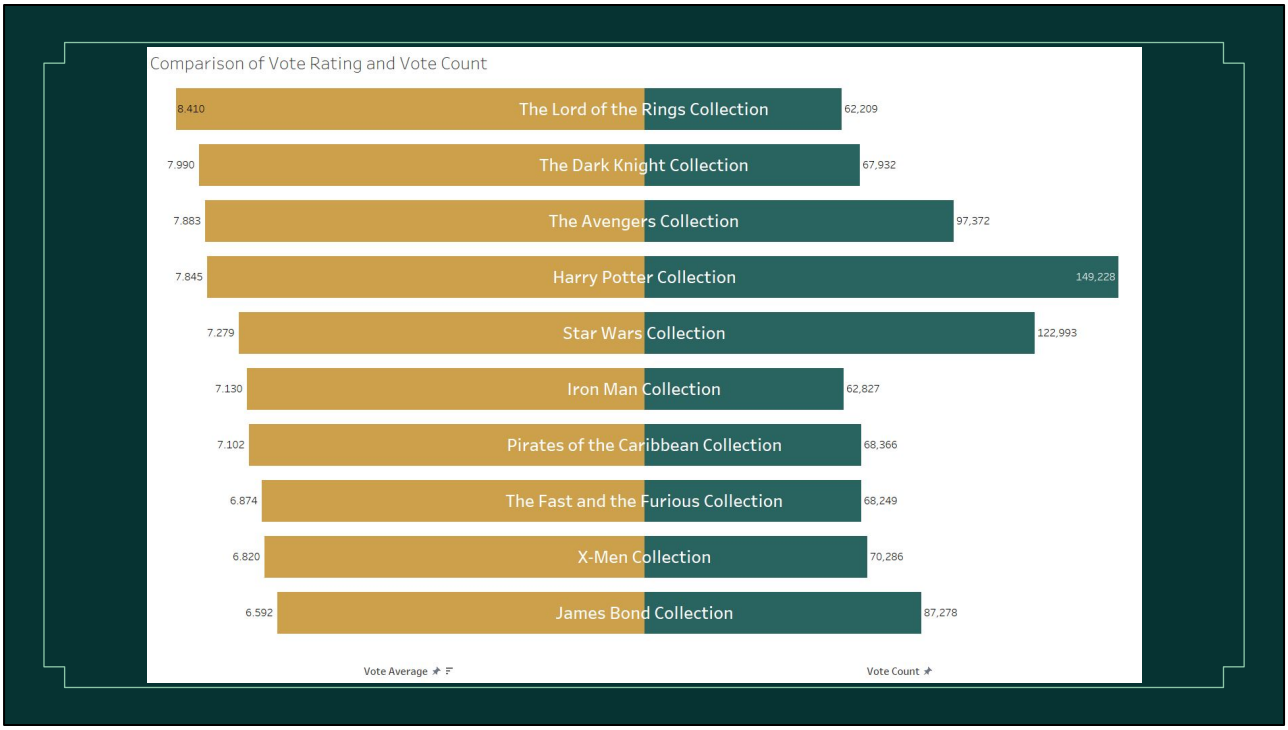


	Popularity	Runtime	Vote Average	Vote Count	Budget	Revenue
Popularity	1	**0.06**	**0.07**	0.16	0.2	0.2
Runtime	0.06	1	0.21	0.21	0.23	0.21
Vote Average	0.07	0.21	1	0.17	**0.14**	**0.14**
Vote Count	0.16	0.21	0.17	1	0.76	0.85
Budget	0.2	0.23	0.14	**0.76**	1	0.84
Revenue	0.2	0.21	0.14	**0.85**	**0.84**	1

With so many variables, I wanted to explore the relationships between them. I started by creating a correlation matrix of selected variables to help guide my analysis. As a reminder, correlation values rank between 0 and 1. The closer to one, the stronger the relationship (note popularity and popularity has a perfect correlation of 1).

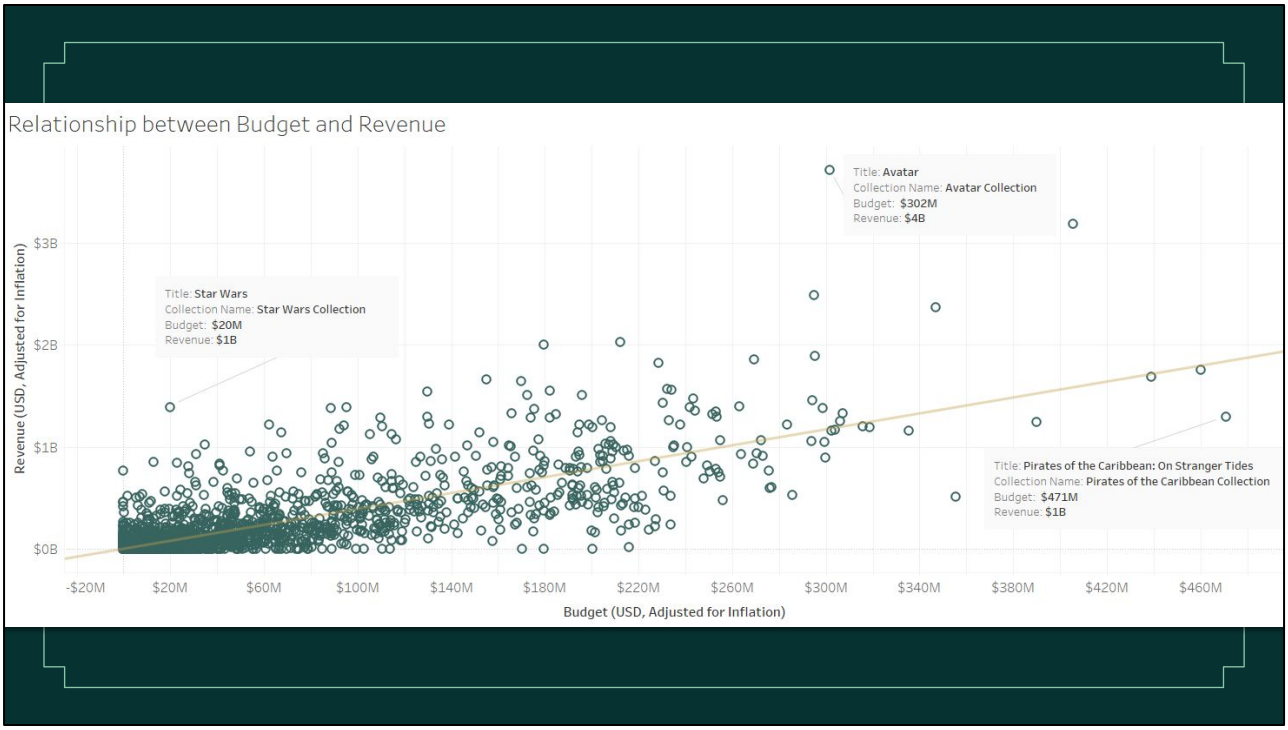
This matrix is colored on a gradient to show relationship strength, ranging from light green to dark green. As you can see, the variables that had the strongest relationships were vote count and budget, vote count and revenue, and budget and revenue.

The weakest relationships were runtime and popularity, vote average and popularity, and (surprisingly), vote average with both budget and revenue. This suggests that TMDB's rankings might not accurately depict consumer sentiment (thus strengthening our use of revenue as an indicator of popularity and measure of success).



Let's first look at the relationship between vote count and vote average (rating). Though not strongly correlated (0.17), it is still useful to paint our analysis with an understanding of TMDB's voting landscape.

The Movie Database reported a total of more than 7 million votes for these films, with an average rating of 5 (on a scale from 1-10). Here's a chart that displays the 10 most-voted for collections and their average rating. As you can see, Harry Potter garnered the *most* votes (with around 150,000), but was edged out by both the Dark Knight and Lord of the Rings Collections, who reported higher *average* ratings.



Next up, let's look at the relationship between budget and revenue. There's a high correlation reported between these two variables (0.84), so it is important to understand that further (remember, we are looking at factors that might help us increase our revenue).

This scatterplot illustrates the relationship between budget and revenue for individual movies. Think of this chart as having four quadrants. The top left quadrant is ideal, showing a high return for relatively low budget. As you can see, there are no films in this quadrant.

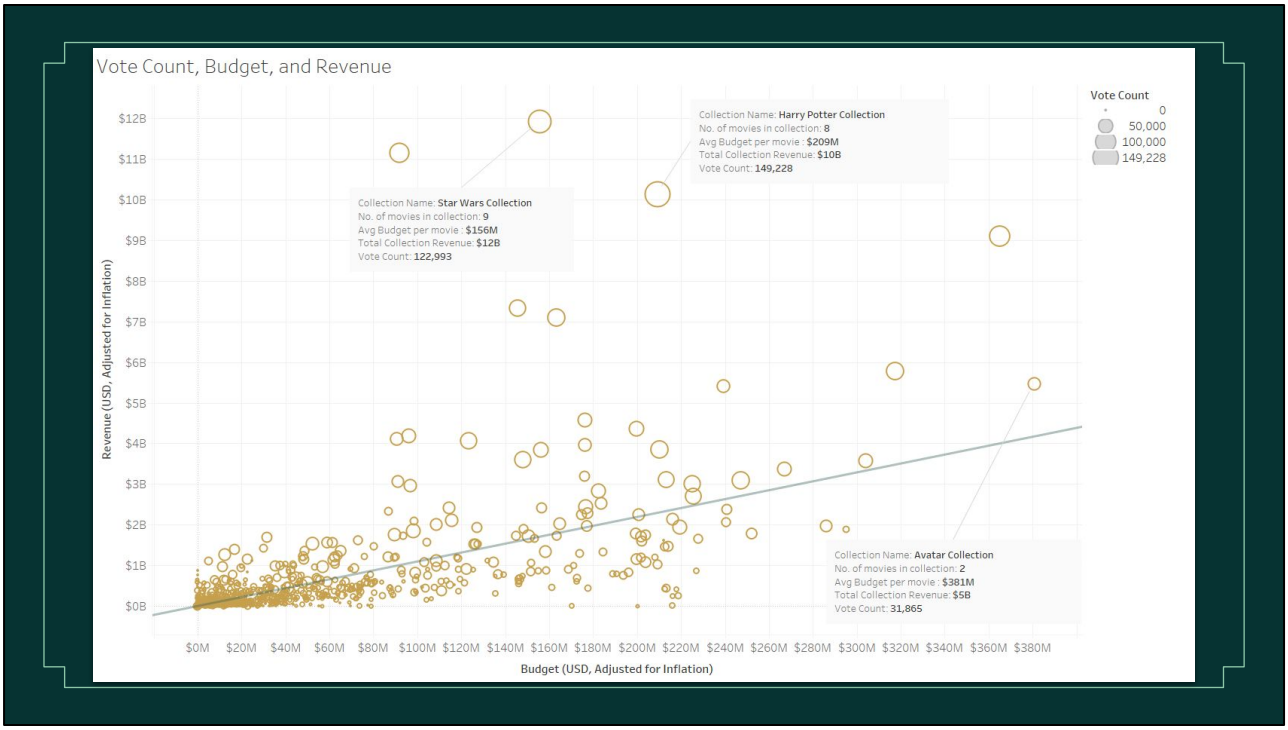
The next optimal quadrant is high return for a high investment, on the upper right side. Avatar is a great example of this.

Next, we move to the lower left quadrant, where the majority of films live. This is low investment but also a low profit. Star Wars stands out in this quadrant, making slightly more than the rest of the pack relative to how much it cost to make.

Finally, the least ideal quadrant is the lower right. This is where films that "flop" live. These are films that have very high budgets but make relatively little money. Pirates of the Caribbean: On Stranger Tides is an example of this one.

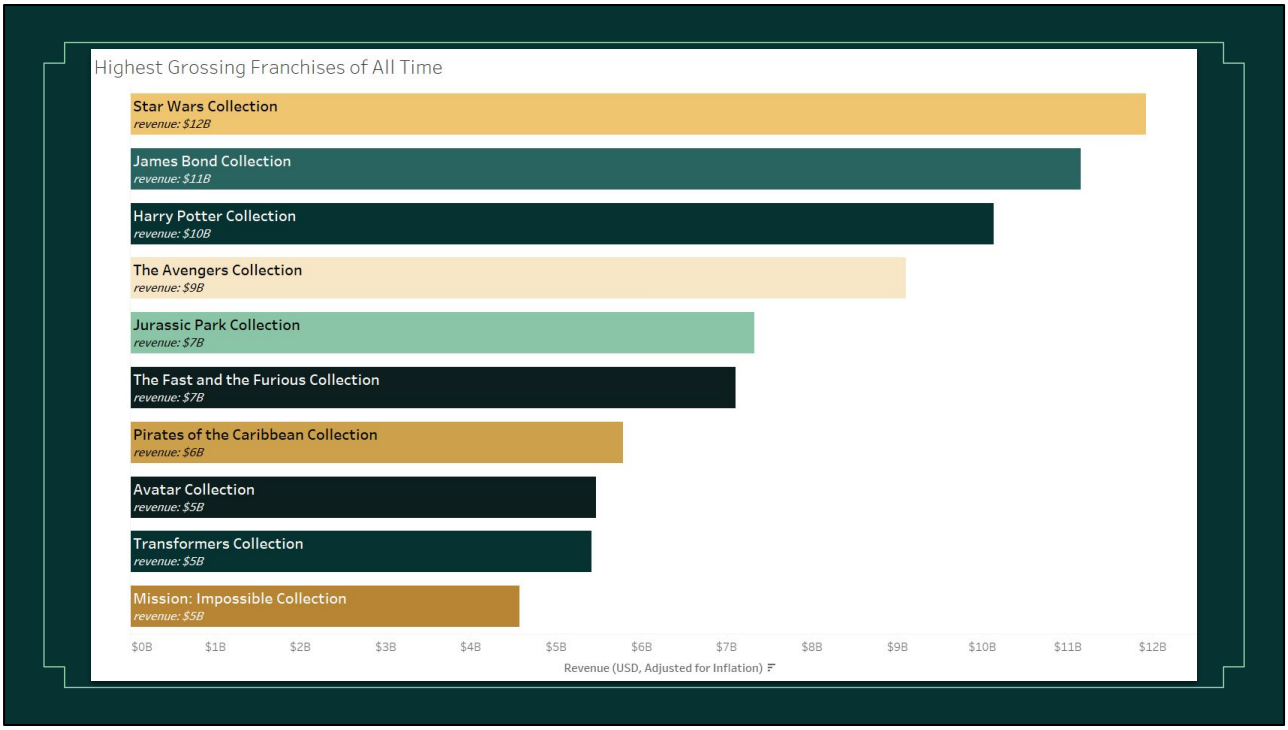
Note that these figures are based on the total revenue generated for the life of the movie. So newer movies like the Avatar sequel, which is currently in this quadrant, might safely expect to move quadrants across its lifespan (it was released in Dec

2022).

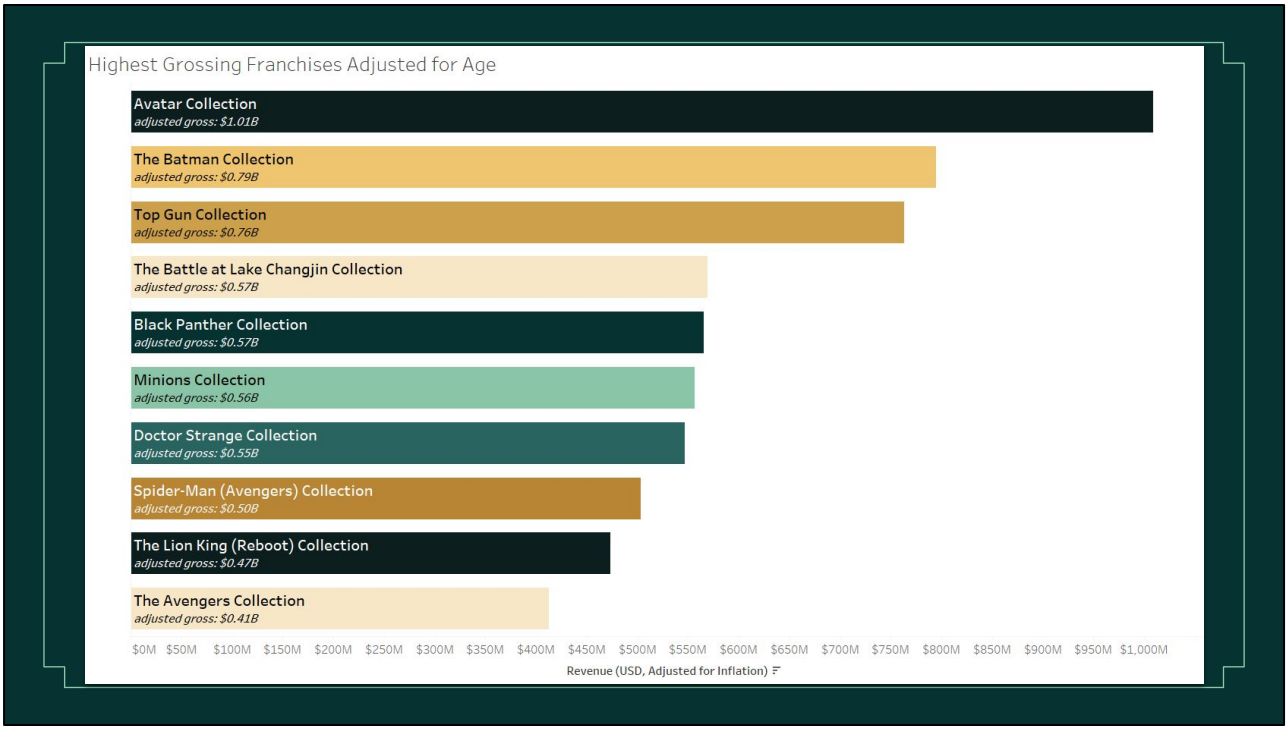


Finally, recall that vote count had strong relationships with both budget and revenue. This could be explained by a variety of things. For instance, movies with high budgets are probably able to spend more on marketing, thus resulting in more viewers (read: tickets bought), and reviewers (vote count).

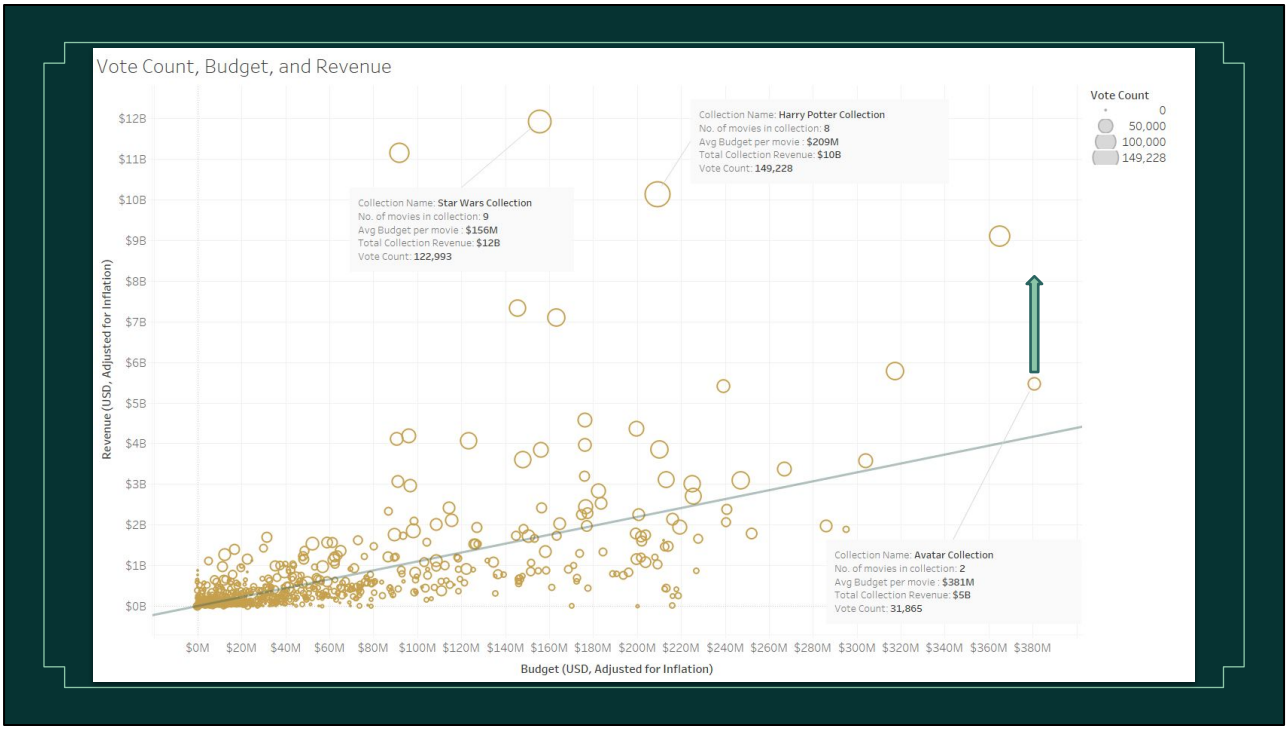
Check out this relationship visualized. This is the same layout as the budget-revenue chart from before, but now each circle represents an entire collection rather than an individual movie. The x-axis depicts the average budget of an individual movie in the collection and the y-axis displays how much the collection made overall. The size of the circle represents vote count. We see Harry Potter over here (which, remember, had the highest vote count of all collections). Relative to the rest of the collections, it had a pretty high per-movie budget and generated considerable revenue, though arguably Star Wars was more successful because it spent less per-movie and generated more revenue.



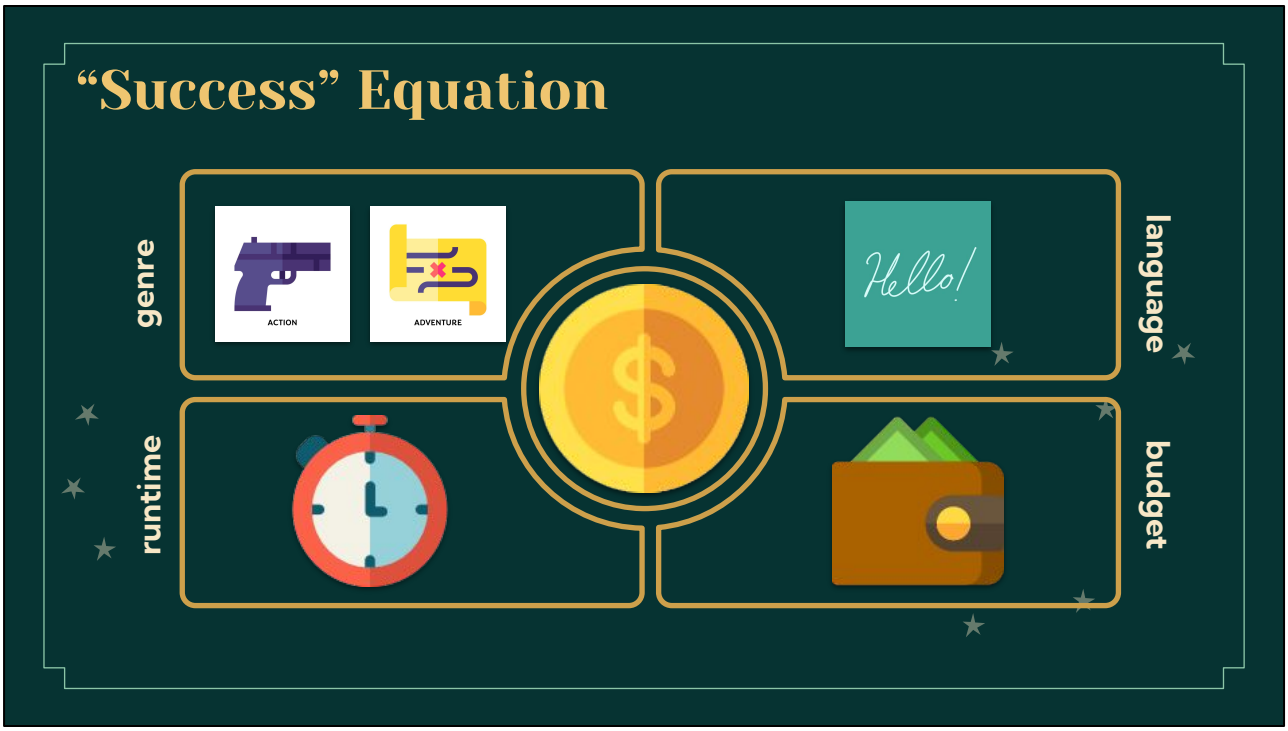
One other consideration for this dataset is that it doesn't account for how long a collection has existed. For instance, Star Wars has generated the most revenue but has also existed since the early 1970s, whereas Avatar only became a collection (that is, the second film was released), in December of 2022. Here's a list of the highest-grossing franchises of all time (see Star Wars at the top)...



…and now here's a list of the highest-grossing franchises accounting for the average age of the collection. Avatar is at the top! Star Wars isn't even on there!

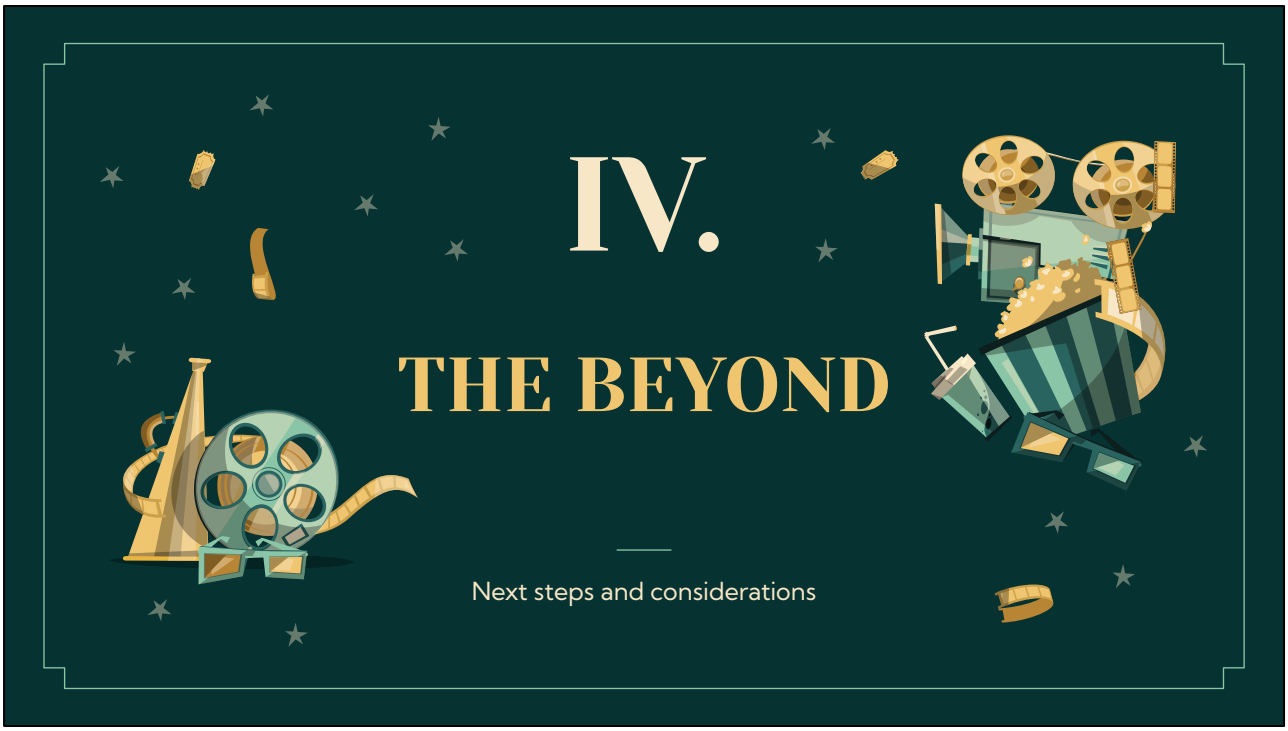


So, going back to our main visualization for a second, while Avatar has the highest per-movie budget, its full earning potential will undoubtedly grow in the future.



Just to drive the point home, lets look at the profile of the ULTIMATE revenue-generating collection (highest ratings across all categories).

- Genre: The genre that generates the MOST revenue overall is Action ($132B). The genre that has the highest AVERAGE revenue is Adventure ($130m per movie)
- Language: English
- Runtime: Matters very little
- Budget: The higher, the better! (As budget increases, so does revenue).



So what have we learned?

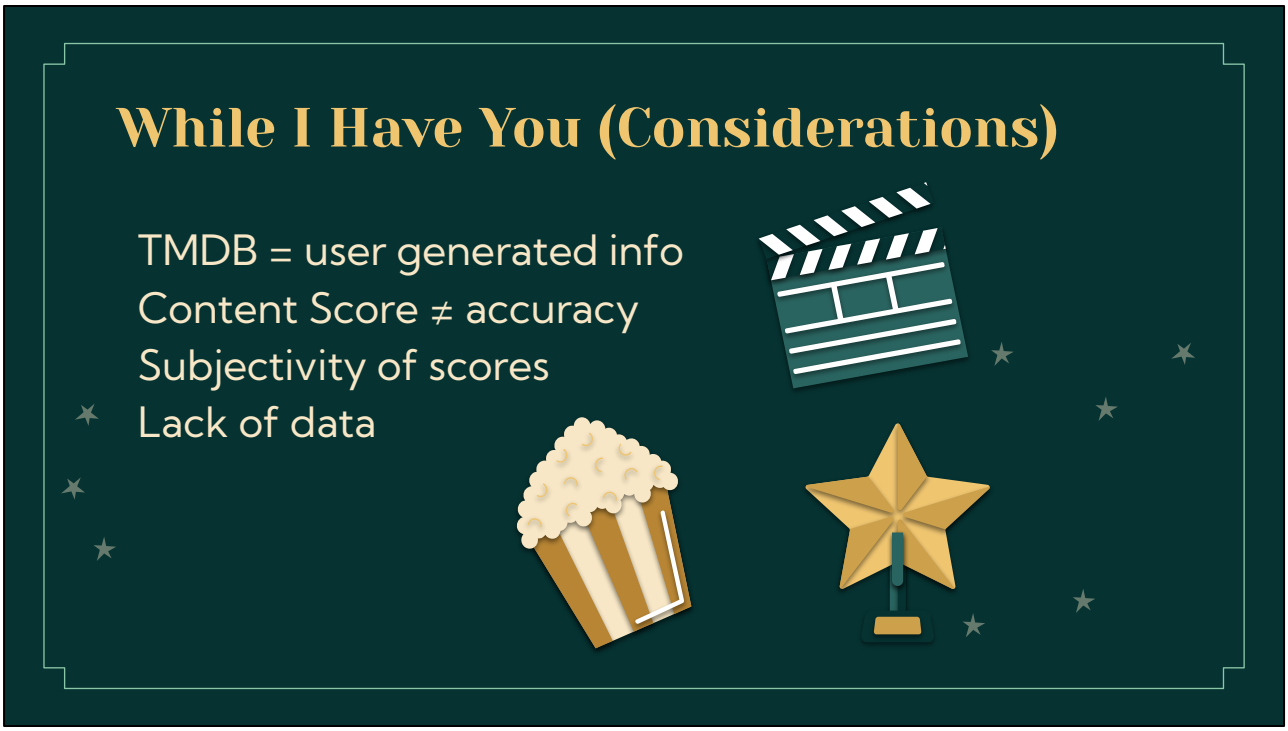
Aside from the obvious, that collections with higher budgets earn more money at the box office, we learned that both high-spending and high-earning films correlate with more engagement on TMDB. The more widely known a franchise becomes, the more people interact with it. The greater the interaction, the greater the opportunity to generate revenue (both from direct viewership and other streams such as merchandising). Therefore, sites like TMDB, IMDB, and Rotten Tomatoes are crucial for franchise development. As we learn more about franchises, it is clear that understanding the consumer (the *viewer*) is crucial to understanding how to make a successful franchise. This leads me to my recommended next steps.

Next Steps

- ✓ Viewership
- ✓ Regression model
- ✓ Other definitions of success
- ✓ Sequels vs originals
- ✓ Comparison to non-franchise films
- ✓ Geographic impact

In addition to exploring viewership on a deeper level, there are some other avenues that will enrich our understanding of franchises. First of all, I'd like to create a regression model to attempt prediction rather than just relationship. We could experiment with other definitions of what makes a franchise successful. We could examine if sequels do better or worse than the original films. We could pull more data to include non-franchise films. Finally, we could explore the role geography plays in box-office success. As movies become more widely available and more creators around the world are able to access the medium, the makeup of successful franchises, and movies in general, has the potential to shift dramatically.

For what it's worth, I've begun to explore these other avenues and will have them available to review soon. I made a genres dashboard if you're interested in that; it's available on my Tableau Public.



There are a few considerations to be had around my analysis. Notably, I used data from The Movie Database, which is entirely user-generated. While every effort is made to ensure accuracy, there are likely to be inaccuracies with reporting (To help this, TMDB has included a “Content Score” for each entry, which depicts the completeness of the information. It doesn’t rate accuracy though.)

Further, subjective fields like rankings likely aren’t representative of the entire moviegoing population.

Importantly, when searching for data for my project, I discovered that there is a notable lack of publicly available, scale-able data for movies. There is a lot of proprietary information that is protected by various studios and streaming platforms. This makes any kind of large-scale, relevant analysis quite difficult.


THAT'S
A WRAP!

ATTRIBUTIONS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Many thanks to **The Movie Database** for the data
Thanks also to **Storyset** for the animations