# **Data source (Last accessed 10/5/2022)**

<https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata>

Background

What can we say about the success of a movie before it is released? Are there certain companies (Pixar?) that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Film aficionados might have different interests. Can we predict which films will be highly rated, whether or not they are a commercial success?

This is a great place to start digging in to those questions, with data on the plot, cast, crew, budget, and revenues of several thousand films.

Data Source Transfer Summary

We (Kaggle) have removed the original version of this dataset per a [DMCA](https://en.wikipedia.org/wiki/Digital_Millennium_Copyright_Act) takedown request from IMDB. In order to minimize the impact, we're replacing it with a similar set of films and data fields from [The Movie Database (TMDb)](https://www.kaggle.com/datasets/tmdb/themoviedb.org) in accordance with [their terms of use](https://www.themoviedb.org/documentation/api/terms-of-use). The bad news is that kernels built on the old dataset will most likely no longer work.

The good news is that:

* You can port your existing kernels over with a bit of editing. [This kernel](https://www.kaggle.com/sohier/getting-imdb-kernels-working-with-tmdb-data/) offers functions and examples for doing so. You can also find [a general introduction to the new format here](https://www.kaggle.com/sohier/tmdb-format-introduction).
* The new dataset contains full credits for both the cast and the crew, rather than just the first three actors.
* Actor and actresses are now listed in the order they appear in the credits. It's unclear what ordering the original dataset used; for the movies I spot checked it didn't line up with either the credits order or IMDB's stars order.
* The revenues appear to be more current. For example, IMDB's figures for Avatar seem to be from 2010 and understate the film's global revenues by over $2 billion.
* Some of the movies that we weren't able to port over (a couple of hundred) were just bad entries. For example, [this IMDB entry](http://www.imdb.com/title/tt5289954/?ref_=fn_t%E2%80%A6) has basically no accurate information at all. It lists Star Wars Episode VII as a documentary.

Data Source Transfer Details

* Several of the new columns contain json. You can save a bit of time by porting the load data functions [from this kernel]().
* Even in simple fields like runtime may not be consistent across versions. For example, previous dataset shows the duration for Avatar's extended cut while TMDB shows the time for the original version.
* There's now a separate file containing the full credits for both the cast and crew.
* All fields are filled out by users so don't expect them to agree on keywords, genres, ratings, or the like.
* Your existing kernels will continue to render normally until they are re-run.
* If you are curious about how this dataset was prepared, the code to access TMDb's API is posted [here](https://gist.github.com/SohierDane/4a84cb96d220fc4791f52562be37968b).

New columns:

* homepage
* id
* original\_title
* overview
* popularity
* production\_companies
* production\_countries
* release\_date
* spoken\_languages
* status
* tagline
* vote\_average

Lost columns:

* actor*1*facebook\_likes
* actor*2*facebook\_likes
* actor*3*facebook\_likes
* aspect\_ratio
* cast*total*facebook\_likes
* color
* content\_rating
* director*facebook*likes
* facenumber*in*poster
* movie*facebook*likes
* movie*imdb*link
* num*critic*for\_reviews
* num*user*for\_reviews

Open Questions About the Data

There are some things we haven't had a chance to confirm about the new dataset. If you have any insights, please let us know in the forums!

* Are the budgets and revenues all in US dollars? Do they consistently show the global revenues?
* This dataset hasn't yet gone through a data quality analysis. Can you find any obvious corrections? For example, in the IMDb version it was necessary to treat values of zero in the budget field as missing. Similar findings would be very helpful to your fellow Kagglers! (It's probably a good idea to keep treating zeros as missing, with the caveat that missing budgets much more likely to have been from small budget films in the first place).

Inspiration

* Can you categorize the films by type, such as animated or not? We don't have explicit labels for this, but it should be possible to build them from the crew's job titles.
* How sharp is the divide between major film studios and the independents? Do those two groups fall naturally out of a clustering analysis or is something more complicated going on?

Acknowledgements

This dataset was generated from [The Movie Database](https://www.kaggle.com/datasets/tmdb/themoviedb.org) API. This product uses the TMDb API but is not endorsed or certified by TMDb.  
Their API also provides access to data on many additional movies, actors and actresses, crew members, and TV shows. You can [try it for yourself here](https://www.themoviedb.org/documentation/api).