# REPORT

The Movie Database (TMDb) is like a big online encyclopedia for movies and TV shows. It was started in 2008 by Travis Bell as a place to share film posters. It began with a donation from the Open Media Database project and has grown a lot since then. Now, it's one of the busiest databases on the internet, used by millions every month. It has info on 543,523 movies, 92,367 TV shows, and a community of about 1,650,750 people. The main idea is to help people learn more about films and shows.

Making a movie is expensive and risky. Some cost over $100 million to make, and they can either do really well or not make much money at all. So, figuring out what makes a movie successful is a big deal for the movie industry. Things like having a famous director or popular actors usually help, but it doesn't always guarantee a hit with the audience. The goal of this project is to use data from TMDb to understand what factors make a movie successful—meaning it makes a lot of money and has a high Tmdb score. By using data science tools, the project aims to uncover the secrets that can help movie companies create blockbuster films.

## Overview of the Dataset

The project begins with an analysis of the data. A clear understanding of the subject as well as of the data, is crucial for drawing relevant conclusions and developing an appropriate model. The data is composed of 24 variables for 5043 films, covering 100 years in 66 countries. There are 52,234 crew members and 54,201 actors.

**Data:**
The dataset contains information about 5000 films grouped into 2 xlsx files:

- tmdb_5000_movies.xlsx : Provides film-specific information such as score, title, release date, genre, revenue etc.
- tmdb_5000_credits.xlsx : Includes information about the actors and crew of each Movie.

The data set contains 24 variables that we can divide into two categories: qualitative variables (17) and quantitative variables (7).

- **Qualitative variables**: Includes 17 variables in the data sets and designates discrete units. It is used to label variables that have no quantitative value and can be found in both csv files.

-In the **tmdb_5000_movies.xlsx** file the qualitative value include the title and the id of the film, the genre of the movie (Sci-Fi, Family, Horror, Comedy, Action,..), the links of the homepage of the film, the language of the film (English, Arabic, Chinese, French,..),the name of the production company (Walt disney, Columbia,..),the status of the film(Released or post production), the tagline of the movie, a small overview as well as keywords describing the film’s plot.

- The **tmdb_5000_credits.xlsx** file Contains 4 variables, two of which are also in the first file and which are the title and the unique ID of the film, the two other variables are in Json format: the casting which contains the names of the main actors with their gender (1 if male 2 if female, 0 if undefined) And the crew, which contains all the members of the film crew, including the name and gender of each crew member, as well as their function and the field in which they work.

The **quantitative variables** are also divided into 2 sub-parts and are only found in the tmdb_5000_movies.csv file:
• **Discrete variables:** Consists of 3 variables in the data set and are numerical data that can only take a certain set of values. The discrete variables are: the number of votes for each film in tmdb, the release date and the duration of the film.
• **Continuous variables:** Consists of 4 variables in the data set and represents measurements or quantities that can take any numerical value. The continuous variables are: film revenues in dollars, film budget in dollars, popularity, as well as average film votes.

In this Project, I have used only tmdb_5000_movies.xlsx data to predict Revenue and Ratings. In future, I will also try to exploit the cast dataset which can give us some more useful insights.

## Data Cleaning:

Unfortunately, the dataset is not perfect, and some problems prevent us from being able to work with it directly.
• The biggest issue with this dataset is the json format (JavaScript Object Notation), it is a syntax for storing and exchanging data between two computers that looks like a dictionary pair (key:value) embedded in a string. Therefore the first task was to browse all the data and replace the json character in string format using the json.loads() method. The columns that create this problem are the genre, keywords, countries of production, production companies, languages spoken in the film file.

• We also need to deal with the null values, five features cause this problem: home page, release date, overview, execution time and slogan. The homepage and the tagline contain the most null values and are not relevant for our study, therefore the entire columns were discarded. This leaves the release date, the run-time and the overview, which contain three, two and one null value respectively. For run-time, the two null values are replaced by the average value of the duration of the films. Then, for the overview, the null value can easily be fixed, we simply put unspecified instead of null in order to have a category for it and thus avoid any problem downstream. Finally, for the release date, we chose to drop this value.

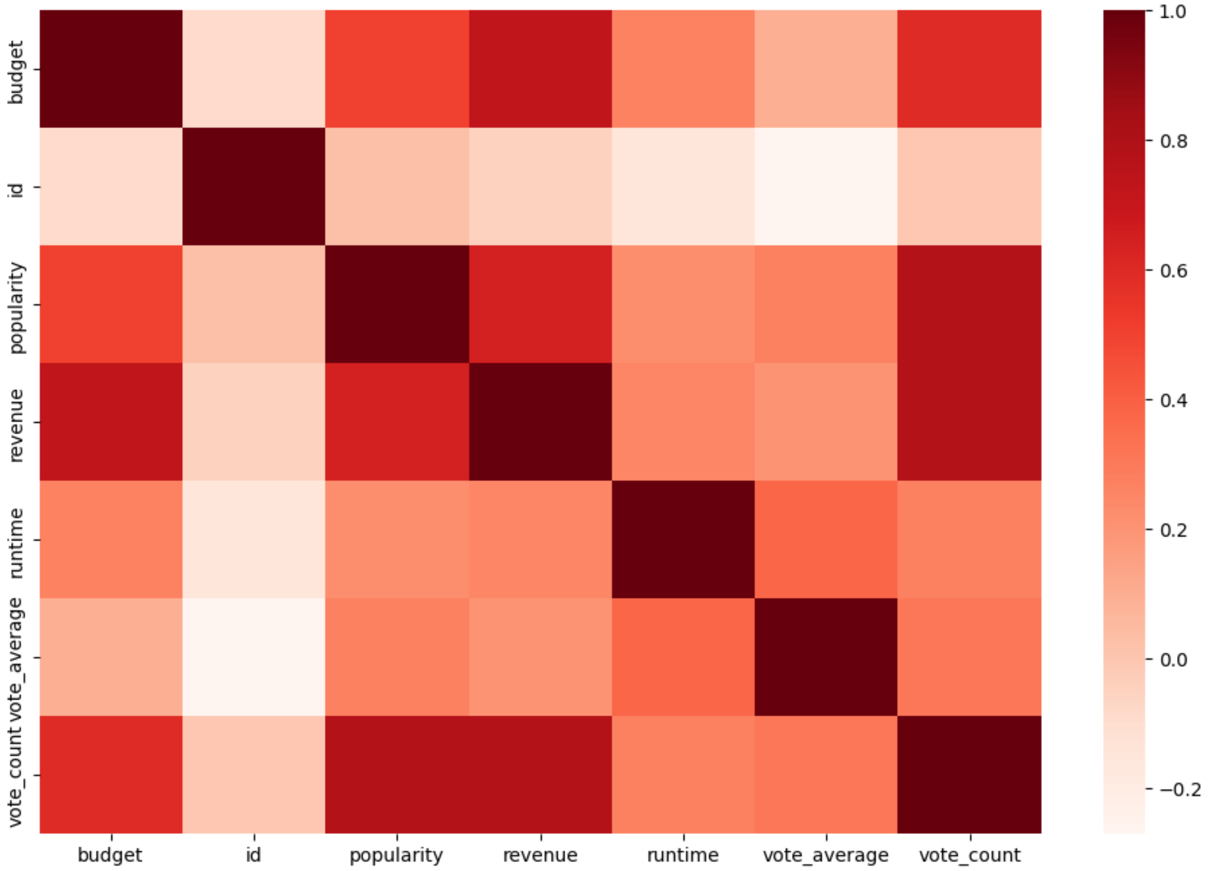
• The last problem is that of zero values. There are five features that cause this problem: budget, revenue, run time, average vote and number of votes. Since our primary objective is to forecast revenue, we are going to drop the zero value for revenue and budgets.
We have decided to drop everything under $1,000, there are cases where the budget

is $5 but in reality it is worth $5 million; however, we opted to drop them instead of replacing them one by one as they only represent a small part of our dataset. Finally, for the other features , we decided to drop them.

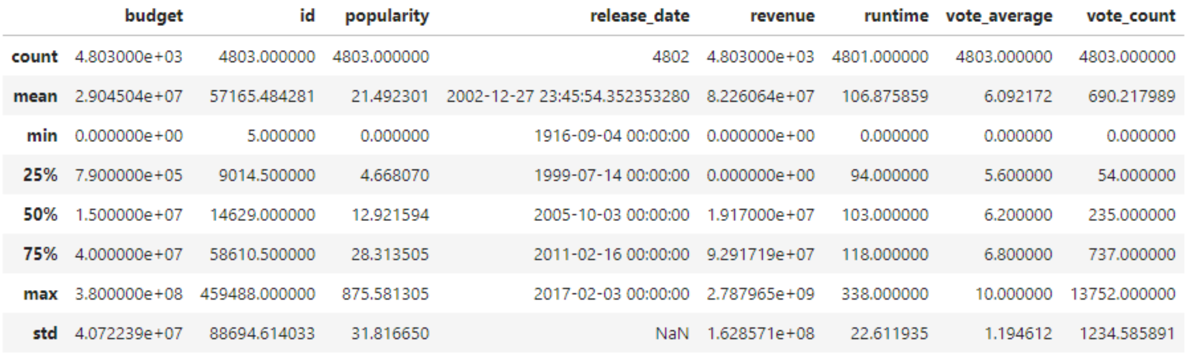
• In addition, there are some movies that have the same title, therefore by adding the year of release, we make the title unique and remedy this problem.

• Also for convenience, in the release date, it is preferable to change the year, month, day format initially in the same column to that of each in a different column.
The data are now ready to be properly observed in order to extract as much information as possible, allowing them to be subjected to statistical tests.

**EDA:**
Before getting to the core of the subject, it is a good idea to first draw the correlation matrix in order to assess the dependency between the different variables. Each cell in the table shows the correlation between two variables.



Statistics summary of movies quantitative features is shown in this figure.



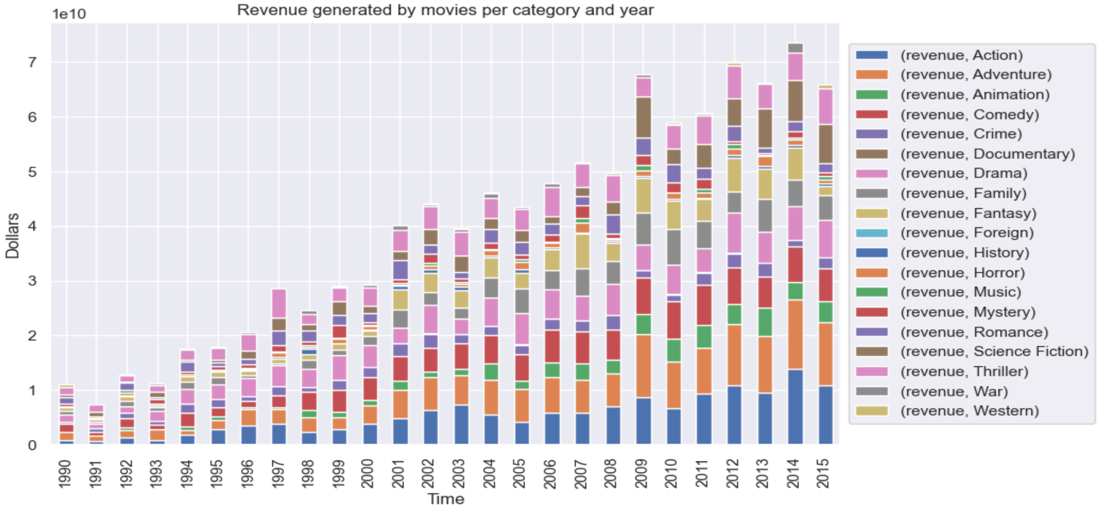
| | budget | id | popularity | release_date | revenue | runtime | vote_average | vote_count |
|---|---|---|---|---|---|---|---|---|
| **count** | 4.803000e+03 | 4803.000000 | 4803.000000 | 4802 | 4.803000e+03 | 4801.000000 | 4803.000000 | 4803.000000 |
| **mean** | 2.904504e+07 | 57165.484281 | 21.492301 | 2002-12-27 23:45:54.352353280 | 8.226064e+07 | 106.875859 | 6.092172 | 690.217989 |
| **min** | 0.000000e+00 | 5.000000 | 0.000000 | 1916-09-04 00:00:00 | 0.000000e+00 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 7.900000e+05 | 9014.500000 | 4.668070 | 1999-07-14 00:00:00 | 0.000000e+00 | 94.000000 | 5.600000 | 54.000000 |
| **50%** | 1.500000e+07 | 14629.000000 | 12.921594 | 2005-10-03 00:00:00 | 1.917000e+07 | 103.000000 | 6.200000 | 235.000000 |
| **75%** | 4.000000e+07 | 58610.500000 | 28.313505 | 2011-02-16 00:00:00 | 9.291719e+07 | 118.000000 | 6.800000 | 737.000000 |
| **max** | 3.800000e+08 | 459488.000000 | 875.581305 | 2017-02-03 00:00:00 | 2.787965e+09 | 338.000000 | 10.000000 | 13752.000000 |
| **std** | 4.072239e+07 | 88694.614033 | 31.816650 | NaN | 1.628571e+08 | 22.611935 | 1.194612 | 1234.585891 |

From the correlation matrix, we observe that the number of votes variable is highly correlated with popularity as well as revenue with a moderate correlation with budget, while the popularity variable is moderately correlated with revenue and slightly less correlated with budget. In addition, the budget and revenue variables are highly correlated. While the run time and average number of votes variables do not seem to have a dependency relationship with the other characteristics.

Then, we begin our data exploration by comparing the genre of each film and its recurrence, we find that the dramatic genre is the most common; it represents 25.18% of all genres, followed by the comedy and thriller and action genres which represent 18.89%, 12.65% and 9.8% respectively. The least recurrent genres are foreign films, westerns and documentaries with a percentage of 2%.

On the other hand, action and adventure films, followed by dramas and thrillers are the most revenue-generating categories, while foreign films, westerns and documentaries are the least revenue-generating categories. Figure below shows the revenues generated for each film category

between 1990 and 2015, where we see a net increase in revenues for each category.



the total revenues generated for each film category during 1990-2015

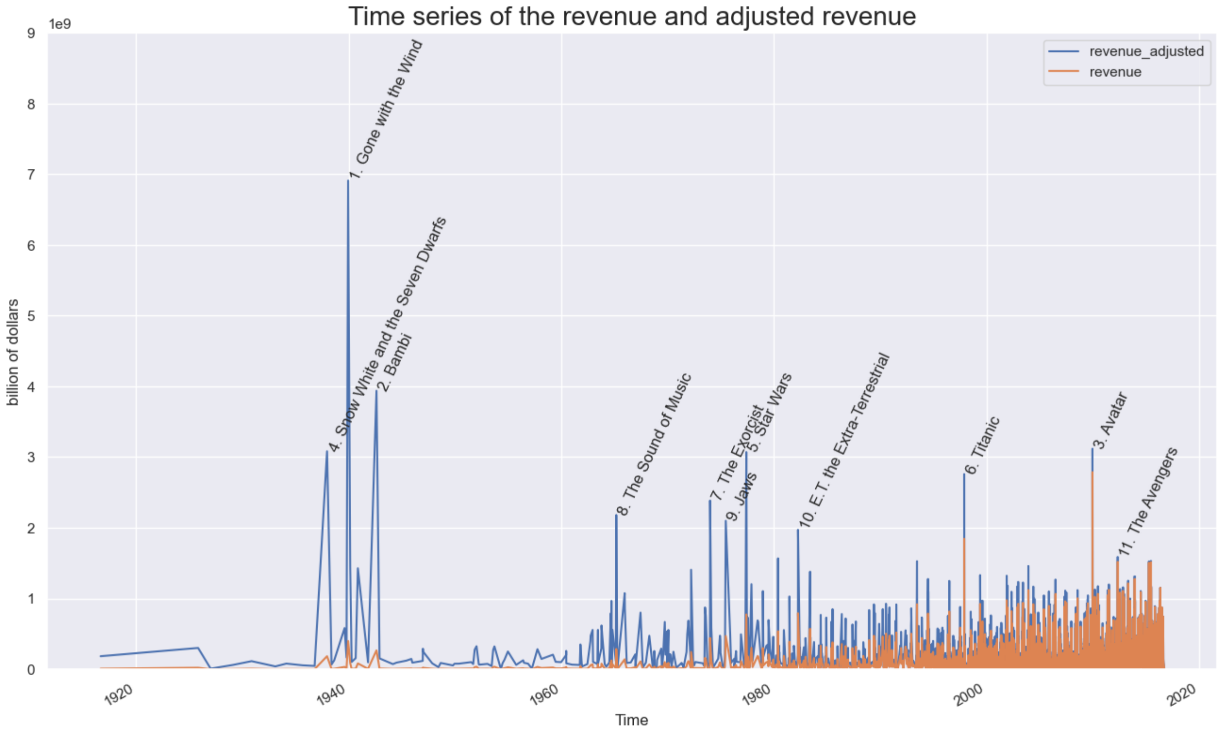
However, this evolution must be nuanced, as the economical phenomenon of inflation must be took into account. The effect of inflation refers to the general and sustainable increase of goods and services prices in the economy. To only take the movie industry as example in 1939, the Americans paid 0.25 $ to watch the blockbuster Gone With the Wind while they paid 5 $ to see Titanic in 1997 and 9 $ for Avatar in 2009. Therefore, in order to really see if the film industry is generating more revenue, it is wise to contextualize the revenues in such a way that the money generated can be more fairly compared, this is done using the inflation increase formula shown below:

Rise in inflation =(CPI ref,year) / (CPI fin,year)

with:

- CPI ref,year: The Consumer Price Index for the reference year
- CPI fin,year: The Consumer Price Index for the final year(here is 2015)

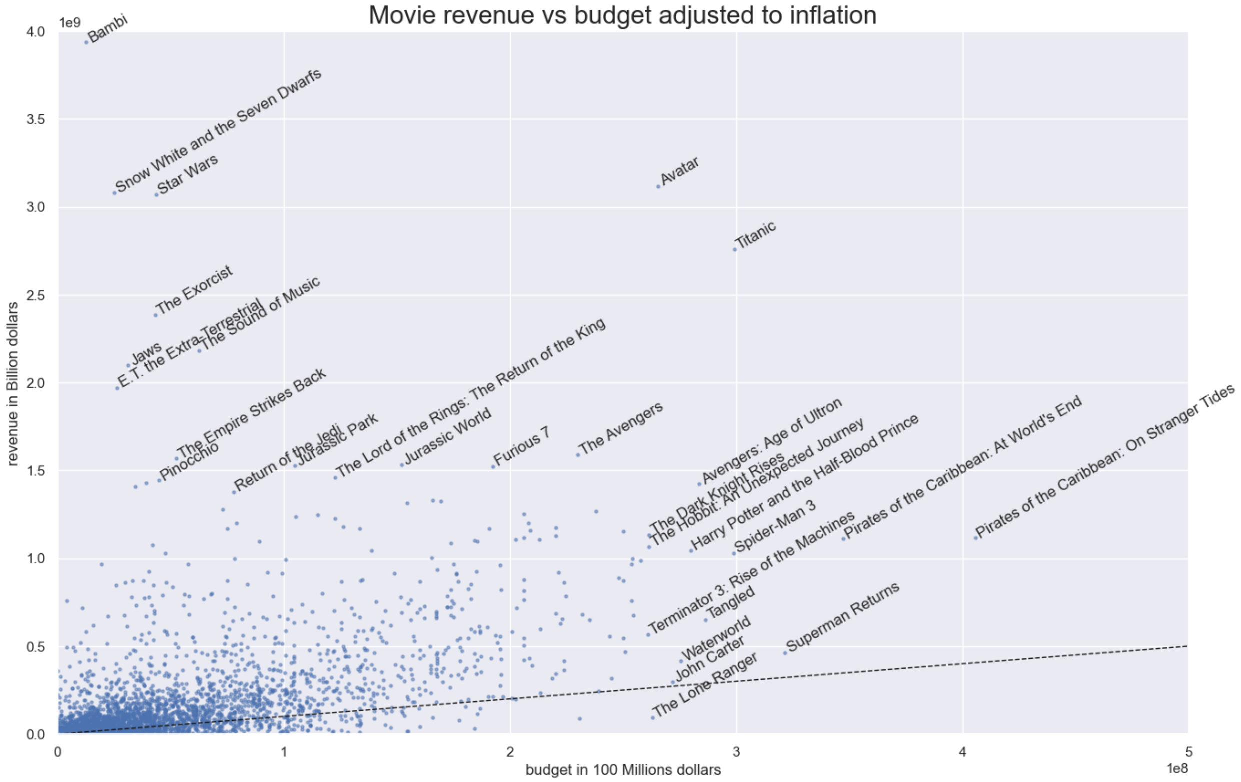
This can be seen clearly in figure below, the orange curve is the curve of the inflation-adjusted revenue while the blue curve is the curve without inflation adjustment, it can be seen that the closer we get to the final time (here 2015) the more the two curves overlap while the further we get from 2020 the more the curves differ. moreover we notice that outliers emerge, as Gone with the wild in 1939 and Star wars V in 1980 or Avatar in 2009.



Time series of the revenue generated during 1990-2015 (the blue curve is adjusted to inflation while the orange is not)

We are surprised to find out that the movie Gone With the Wind (1939) became the highest grossing movie with almost 7 billion dollars followed by the more surprising Walt Disney movie Bambi (1942) with 4 billions while the known Avatar is (only) at 3 billions. So how can we explain these differences ? Well actually what is not shown in this graph is that the re-releases of the movie. Gone with the Wind was a big hit at the time of its release but only made 500 millions in current dollars, it was then re-released more than 8 times which explains its success. The takeaway message here is that movies don't only generate money at their release but continue also to do so years after that.

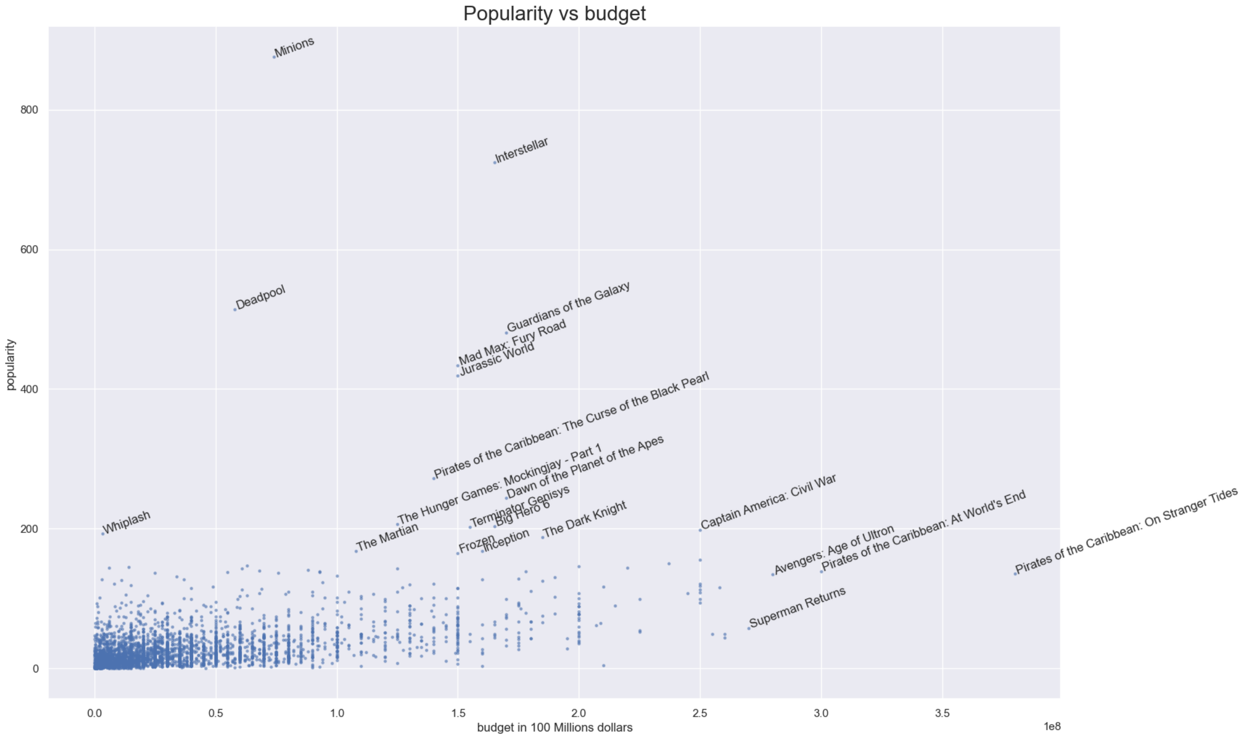
Furthermore, it would be interesting to consider the revenues of the films according to the budget with adjustment for inflation (below figure), this will allow us to see the net result of each film. If the income is higher than the budget, we speak of a profit, whereas if it is lower than the budget, we speak of a deficit. The dashed line represents the case where the income is equal to the budget, so all the films below this curve report a loss, and all the films above this curve report a gain, so the further the film is from this curve, the greater its gain/loss.



Film revenues versus budget with adjustment for inflation (the dashed line represents the case where the budget is equal to the revenues)

In addition, it can be seen from the graph that the majority of films generate a profitable result (above the dashed line) and in general, the higher the budget, the higher the revenue, in accordance with the correlation matrix (1st figure). Nevertheless, there are films that have a very big budget such as The lone ranger, for example, but are suffering a loss and are far from breaking even, furthermore there are many films that have a small budget but generate huge revenues such as the films of bambi, Snow white and the seven dwarves, these latter can be considered as outliers due to their low numbers compared to the rest of the dataset.

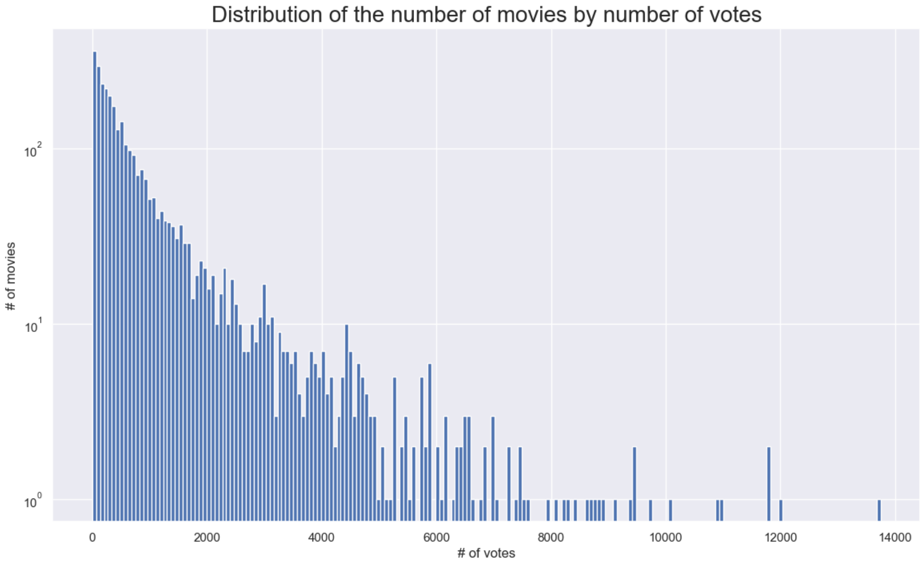
Furthermore, an interesting aspect to study would be the relationship between a film’s popularity with the public and its respective budget. The popularity is directly calculated by TMDB by taking into account several factors which are: The number of votes for the day, the number of views for the day, the number of users who marked it as a "favorite" for the day, the number of users who added it to their watchlist for the day, the release date, the total number of votes, the score of the previous days.



Film popularity (TMDB) versus budget

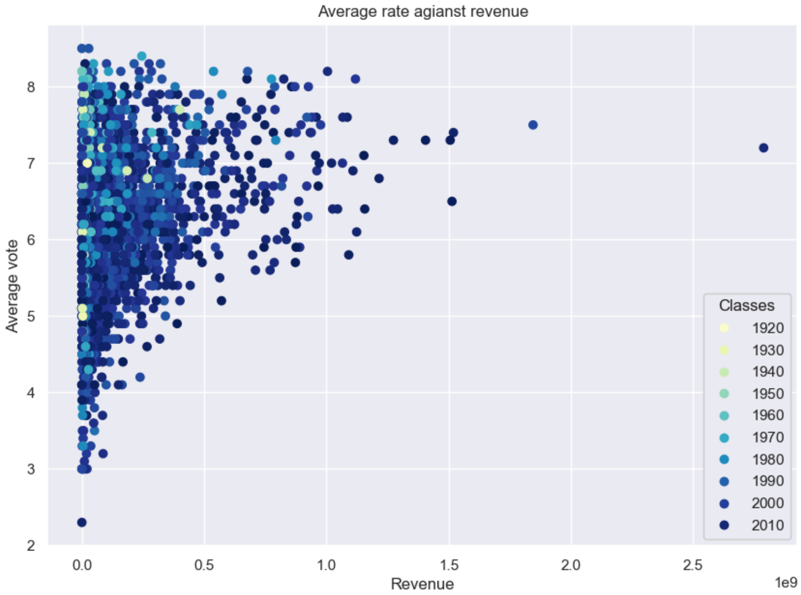
On one hand, we can see on the graph that the majority of films are in the bottom left corner (low budget, low popularity), on the other hand some films have a medium budget but have had a great success with the audience as is the case for The Minions orInterstellar, while some films have a huge budget but have not been very successful with consumers as is the case for *Pirate Of The Carribean : On Stranger Tides or Superman Returns*, therefore all these films can be considered as outliers. Also, another important thing to see is that these outliers are different from the figure of revenue vs. budget, which shows that a film that is very popular with consumers is not necessarily the one that generates the most money, contrary to what one might think.

**Relation between average rating and Revenue :** During the EDA we wanted to find if there is any relation between the average rating given by viewers and the revenues of the movie, but this relation is affected by another criteria which is the number of rating shown in the figure below :



Distribution of the number of movies by the number of votes

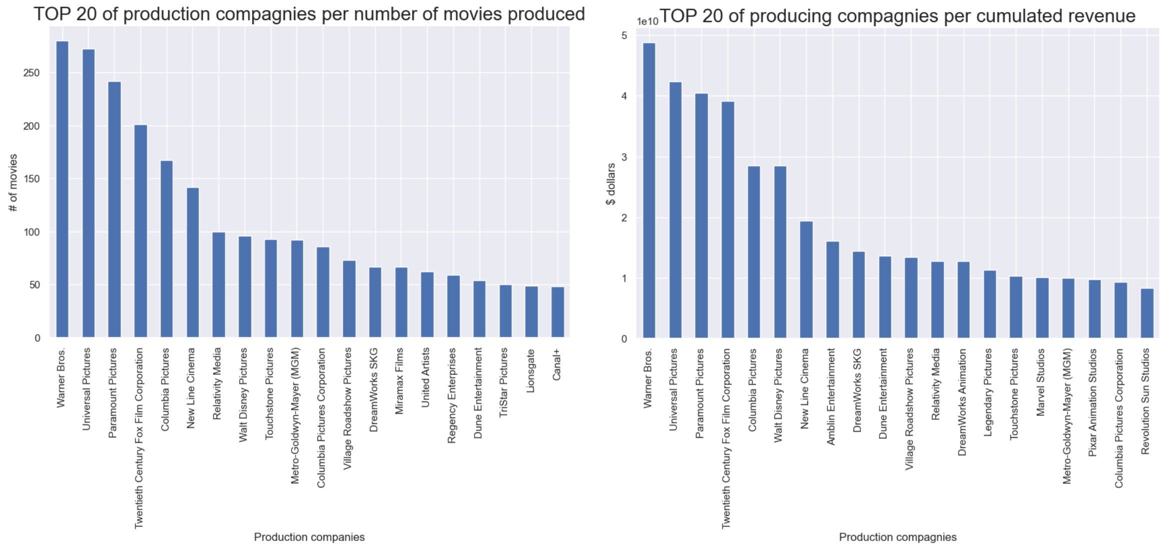
As we can see in the figures that there are about 160 movies have no rating. The next step consists of dropping the movies without any viewers rating and analysing the relation between revenues and rating grouped by decade as we can see in the figure below:



Average rate vs revenue

It can be seen that for the majority of movies the revenues increases when the rating increase and we can observe one outlier with rating of 7.2 and a revenue about 2.8 billion dollars.

**Production companies :** In the data set we found 3539 production companies, but there is a big number of these companies that have only one movie, so we chose to work with the top 20 companies based on the number of movies produced, these 20 companies are kept for the model derivation. The figure below shows the number of movies produced and the cumulated revenues of the 20 top companies.



Top 20 companies by number of movies produced and cumulated revenue

## Data Preprocessing for Model formation

To predict movies revenues we need first to make some changes on the initial data set so it can be used by the chosen algorithms, after data cleaning we did further data preprocessing such as:

**Creating dummy variables:** Since our data set cannot be used by our algorithms, we use one-hot encoding method and create the different categorical variables from the features original language, country of production,production companies, actors, crew names and genre.

**Unused features:** Before starting the revenue and rating prediction we drop the features revenue, number of votes, Rating average and popularity because in the first place we will get these information after the release of the movie and what we are trying to do here is to predict whether the movie is successful or not also if a movie have an important revenue it means that it’s popular which leads to an increase in the number of votes and rating.

Naturally, we will divide our data into two sub-samples : one is used to train the model and the other to test the model using prediction.

## Prediction Model:

We will try to derive a model to predict movies revenues and movies ratings. We will be working with the films dataset without any features about crew members or cast.

We will try to fit 3 different models : Ordinary Least Squares (**OLS**) regression, **Ridge** regression and **XGboost**, and try to find the most suitable by evaluating their performance.
Regression (OLS) treats all variables equally, whereas the ridge regression model is a regression model that has regularization and can rescale its weights by a hyperparameter alpha. We tune alpha to have the best score for this model. This said, we will use iterative feature reduction based on statistical significance when fitted on the data with crew and cast by the OLS model as this will reduce the number of parameters the model have to estimate and gauge the importance of the features.
The XGboost (eXtreme Gradient Boosting) model is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Indeed, such as random forest XGboost leverage the power of boosting but with the difference that now errors are minimized by gradient descent. It achieves superior results by mixing system optimization like parallelization and tree pruning with algorithmic enhancement such as regularization and cross-validation.
The metric we will be using for evaluating the different models will be the adjusted r squared as it is a good evaluation for regression models, it compares the descriptive power of the models. This metric compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. The baseline will be predicting all movies revenues as the mean of revenues in the training set. The best score you can get is 1 and the worst is minus infinity.

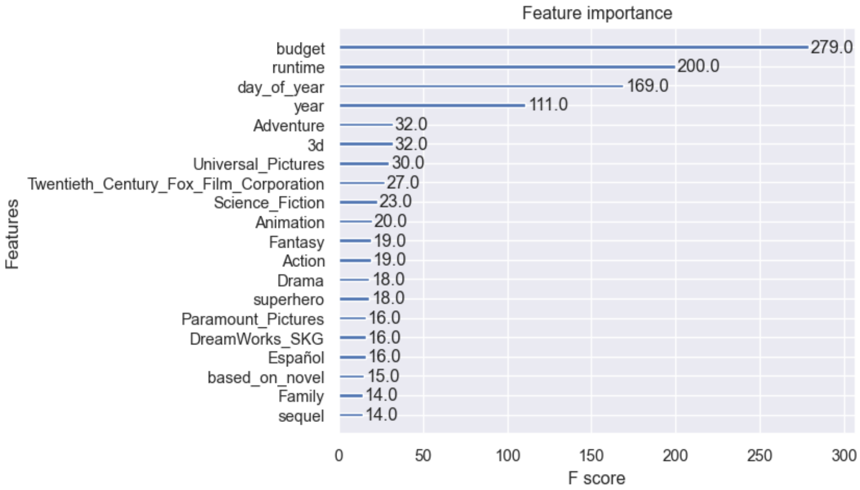
## Predictions:

**Revenue prediction:**
Our baseline will be the mean predictor with a score of -0.005.

**OLS:** -0.0514
**Ridge regression:** 0.502
**XGboost:** 0.541

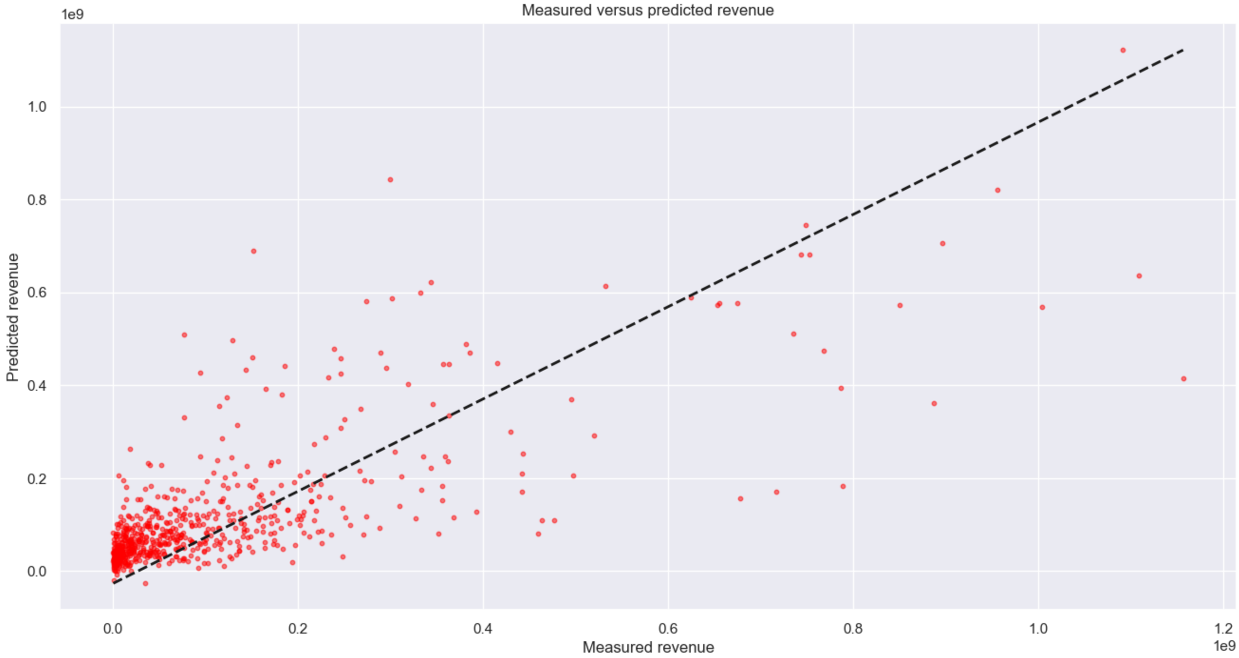
The OLS model performs very poorly on this kind of data as it has too many features. In this framework we can conclude that XGboost outperforms both Ridge Regression and OLS.

For the data without cast, we can see by the figure below, the most important features: Budget, runtime, day of the year and year, Universal pictures and so on. Thus, the budget has a big influence on the success of a movie and its revenue which is logical. The runtime of a movie plays a role too, as if it is a short movie it wouldn't be so interesting and if it's too long it might get boring. We observe that we have some of the big production companies too, Universal Pictures, Twentieth Century Fox Film Corporation, Walt Disney Pictures and Warner Bros, as their movies tend to be most successful.



Most important features by XGboost, the model retained 60 features.

We can see here our predicted values as compared to the Measured values:



Measured versus predicted revenue

**Rating Prediction:**

Our baseline will be the mean predictor with a score of -0.001.

Predictions of our models:

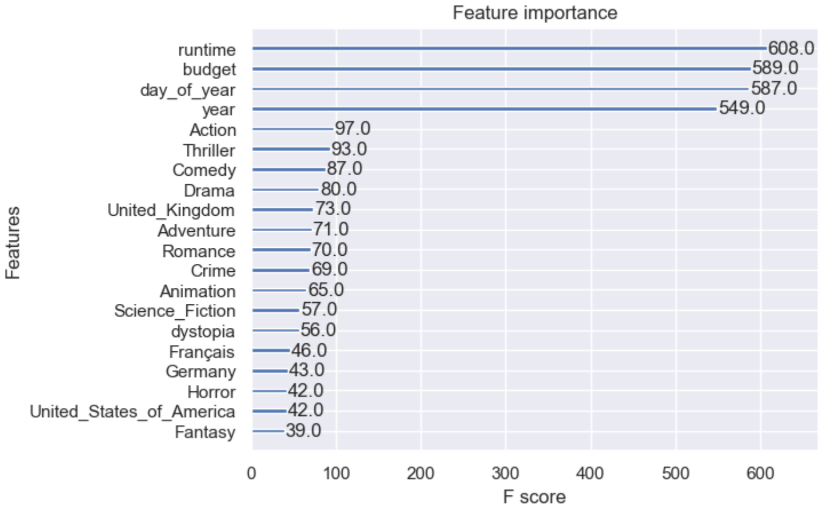
**OLS:** -56.0

**Ridge Regression:** 0.221

**XGboost:** 0.288

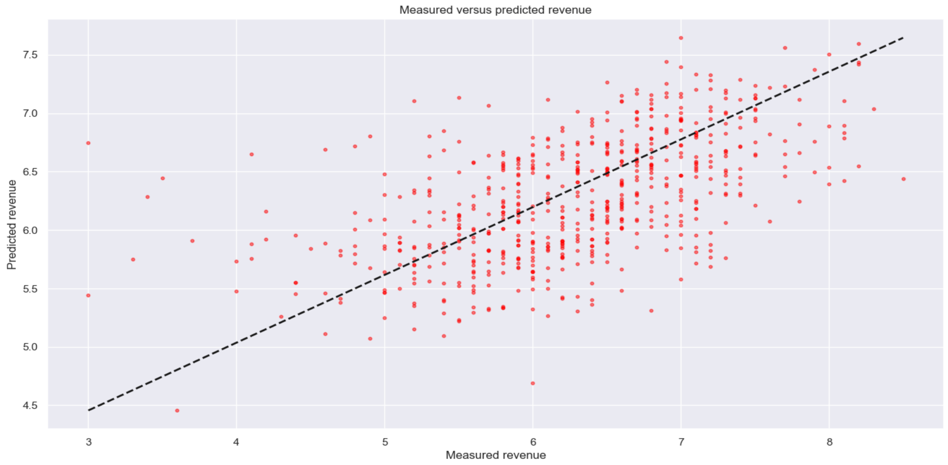
For predicting the ratings, the best model to use is the XGboost model. We can see that it has the best score. Again, the OLS model is performing poorly and this is expected since it is not suited for this kind of regression problems. The ridge regression's score is better but it's outclassed by the baseline so we don't consider it.

For the Rating predictions, we can see by Figure below the most important features: runtime, day of year, year ,budget, Action, Comedy, Thriller... And we find Universal pictures too. Again, we notice the importance of the budget on the ratings. However, the most important feature is the runtime now. We are trying to predict the ratings, so it is only normal that the runtime will be the most significant feature as explained before a long movie can get boring and a short one can be vain. The day of year significance can be inferred by social trends and internet hype over some movies, for example movies during the holidays are quite similar and therefore their reviews can be predicted more easily. The presence of many different genres in the figure can be explained by the fact that each genre has its own fanbase and therefore their own distinct reviewing behavior.



Most important features by XGboost, it retained 73 features.

This graph shows the measured vs predicted ratings



Measured versus predicted rating

## Conclusions

Working on the TMDb movies dataset gave us the opportunity to conduct studies and analysis on real life data about movies. From data cleaning, Exploratory Data Analysis to model derivation we investigated different aspects of movie production and success. We were faced with raw unprocessed data, yet derived a way to process and clean it, and were able to visualize the relation between the movies characteristics and what measures their success and in the process we learned about what makes a movie successful. For instance we analysed how important the budget is for deciding the movie success, obviously this comes as no surprise but here we were able to gauge its importance via statistical metrics. We experimented with ML models to find the best suitable one for our data and goals. Our goal from the beginning of this project was to derive a good enough model to predict the success of a movie beforehand.

In future, I will try to exploit the cast data in **tmdb_5000_credits.xlsx,** which can give us some more useful features to predict the revenues and ratings with high accuracy.