Camera, Lights, Action!

Foundations in Data Science - Group 4 Term Project

Virginia Massignan
Oleksandr (Alex) Rud
Joshua Au-Yeung
Nahid Bhoja

Analyze the relationship between budget, revenue, and popularity to determine factors that contribute to a movie's success.

We can analyse our dataset to try to find factors that contribute to whether the movie was a success. An objective measure of whether a movie was a success or not is whether the movie made the producers money. If the movie made money, there's a high chance that more follow up movies will follow and that the movie may spawn a franchise. Although popularity and ratings can indicate a film success, if the film lost money, its unlikely that there will be follow up movies or franchise merchandise.

So how do we know if a movie made money? We can look for the profit of a film. If the movie brought in a profit (revenue – budget is a positive value), then that indicates the film made some money. But profit alone is not a great indicator. If a movie just made a little bit of profit as a percentage of it's budget, then the movie could still be deemed to be a failure. For example, if a movie budget cost $100 million, and the movie brought in $101 million revenue, the movie is likely to be deemed a failure. But if a movie cost $100,000 and brought in $1 million, then the movie is deemed a resounding success. So a better measure of whether a film made money is the resulting profit margin of the film.

We can define profit margin as the following.

Profit margin = (revenue – budget) / budget x 100%

The higher the profit margin of the film, the greater the movie success is. If the movie had a negative profit margin, then we can easily claim that the movie was not a success. Taking an initial look at the first five films of our dataset and running a few minor calculations, we can see the profit margin of the first 5 films in our dataset.

	Title	**Revenue**	**Budget**	**Profit**	**Profit Margin**
0	Inception	825,532,764	160,000,000	665,532,764	4.159580
1	Interstellar	701,729,206	165,000,000	536,729,206	3.252904
2	The Dark Knight	1,004,558,444	185,000,000	819,558,444	4.430046
3	Avatar	2,923,706,026	237,000,000	2,686,706,026	11.336312
4	The Avengers	1,518,815,515	220,000,000	1,298,815,515	5.903707

These 5 movies by coincidence happen to be huge blockbuster films. They have profit in the hundreds of millions. Avatar by common knowledge is the highest grossing revenue film of all time. It has a profit margin of 11. So this is likely amongst the top profit margin that we'll find.

After cleaning up the data a little to remove movies that haven't yet been released (since we only care about historical data), we take a look at some of the summary statistics of our films:

Column	**count**	**unique**	**top**	**freq**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
title	1,101,128	947,634	Home	152	NaN	NaN	NaN	NaN	NaN	NaN	NaN
revenue	1,101,139	NaN	NaN	NaN	704,848.5	18,026,270	-12	0	0	0	3,000,000,000
budget	1,101,139	NaN	NaN	NaN	267,276.4	4,963,690	0	0	0	0	1,000,000,000
vote_average	1,101,139	NaN	NaN	NaN	1.938	3.048	0	0	0	5	10

popularity	1,101,139	NaN	NaN	NaN	1.248	7.703	0	0.6	0.6	0.873	2,994.357
profit	1,101,139	NaN	NaN	NaN	437,572.1	15,168,740	-1,000,000,000	0	0	0	2,780,000,000
profit_margin	61,778	NaN	NaN	NaN	∞	NaN	-2	-1	-1	0	∞

Its clear that some of this data has issues that need fixing up. For example, the max profit margin somehow reaches infinity.

The problem with IMDB is that anyone can enter in information. People can say that their movie made 3 billion without needing any sources to back it up. So although there may be true information in the database, there is going to be a lot of garbage too.

So we have some bad data. We don't know exactly what films in our dataset are bad without looking individually at films – an unfeasible task, but what we can do is set a list of criteria based on domain knowledge to weed out films that clearly contain incorrect properties.

I set movie threshold criteria to be the following:

revenue_threshold = 3_000_000_000 # 3 billion. I know that Avatar is highest grossing of all time, so that's our max value

budget_threshold = 500_000_000   # 500 million. I know that Avatar 2 and Star Wars Force Awakens are the most expensive movies ever made, so we can use that as our cap

profit_threshold = 3_000_000_000   # 3 billion. Using Avatar 2 revenue upper bound. Naive figure, but need to pick something

Here's our resulting data after cleaning up.

Column	**count**	**unique**	**top**	**freq**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
title	11,075	10,762	Godzilla	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
revenue	11,075	NaN	NaN	NaN	62,075,530	156,289,300	1,015	1,000,000	10,000,000	50,697,340	3,000,000,000
budget	11,075	NaN	NaN	NaN	21,701,540	36,741,750	4	1,028,882	7,520,000	25,000,000	460,000,000
profit	11,075	NaN	NaN	NaN	40,373,990	131,825,600	-199,546,000	-801,546	1,300,000	28,000,000	2,780,000,000

profit_margin	11,075	NaN	NaN	NaN	7.3189	52.1582	-0.9999	-0.2676	0.8533	3.0000	999.9474
vote_average	11,075	NaN	NaN	NaN	5.8393	2.2232	0	5.5685	6.34	7	10
popularity	11,075	NaN	NaN	NaN	15.101	17.806	0	3.5	11.824	19.2705	241.285

We fixed up our data a little. There's likely still some garbage mixed in there, but there's only so much cleaning you can do before you're removing potentially good data as well that contains outliers.

Our mean profit margin for a film is 7.3. This is actually quite surprising to me. I personally had no idea whether movies in general made money or not. This metric implies that movies in general actually have a positive return on their investment. The profit margin standard deviation is 52 though, so there seem to be a lot of variation amongst results.

Another surprising result is that the mean budget of the films is $21 million. The dataset is supposed to be inclusive of films regardless their size. That means it should incorporate small budget indi films. Short and indi films are much more numerous than large budget films simply due to their ease of creation. Those rarely have a budget of even $1 million. Likely this means that small indi films and short films didn't list their budget in the dataset. Our resulting data possibly omitted almost all short and indi films.

We should note that the data fed into this model is still quite unreliable. If I look at the top 10 movies for revenue and budget for example, I can see clear discrepancies where there is obviously incorrect data.

Our standard deviation for budget and revenue is larger than the mean of budget and revenue. This provides some illogical implications. It suggests that its possible for our films to have negative budget and negative revenue. This is logically impossible. This means our budget and revenue data are highly positively skewed since our data should have a lower bound of 0 on the budget and revenue.

Lets take a look at the top 10 highest revenue and budget films of our cleaned dataset.

Top 10 highest revenue films:

Title	**Revenue**	**Budget**	**Profit**	**Profit Margin**	**Vote Average**	**Popularity**
TikTok Rizz Party	3,000,000,000	250,000,000	2,750,000,000	11.00%	10.00	0.00

Bee Movie	2,930,000,000	150,000,000	2,780,000,000	18.53%	0.00	1.40
Avatar	2,923,706,026	237,000,000	2,686,706,026	11.34%	7.57	79.93
Avengers: Endgame	2,800,000,000	356,000,000	2,444,000,000	6.87%	8.26	91.76
Avatar: The Way of Water	2,320,250,281	460,000,000	1,860,250,281	4.04%	7.65	241.29
Titanic	2,264,162,353	200,000,000	2,064,162,353	10.32%	7.90	102.35
Star Wars: The Force Awakens	2,068,223,624	245,000,000	1,823,223,624	7.44%	7.29	66.77
Avengers: Infinity War	2,052,415,039	300,000,000	1,752,415,039	5.84%	8.26	154.34
Spider-Man: No Way Home	1,921,847,111	200,000,000	1,721,847,111	8.61%	7.99	186.07
Jurassic World	1,671,537,444	150,000,000	1,521,537,444	10.14%	6.68	54.09

Top 10 highest budget films:

Title	**Revenue**	**Budget**	**Profit**	**Profit Margin**	**Vote Average**	**Popularity**
Avatar: The Way of Water	2,320,250,281	460,000,000	1,860,250,281	4.04%	7.65	241.29
Lost in the Stars	334,039,200	417,549,000	-83,509,800	-20.00%	6.33	23.73
Pirates of the Caribbean: On Stranger Tides	1,045,713,802	379,000,000	666,713,802	1.76%	6.54	79.19
Avengers: Age of Ultron	1,405,403,694	365,000,000	1,040,403,694	2.85%	7.28	96.57
Avengers: Endgame	2,800,000,000	356,000,000	2,444,000,000	6.87%	8.26	91.76

Avengers: Infinity War	2,052,415,039	300,000,000	1,752,415,039	5.84%	8.26	154.34
Pirates of the Caribbean: At World's End	961,000,000	300,000,000	661,000,000	2.20%	7.24	81.88
Justice League	657,926,987	300,000,000	357,926,987	1.19%	6.10	91.09
The Lego Movie 3: A Powerpuff Adventure	400,000,000	300,000,000	100,000,000	0.33%	0.00	0.00
Superman Returns	391,081,192	270,000,000	121,081,192	0.45%	5.74	31.19

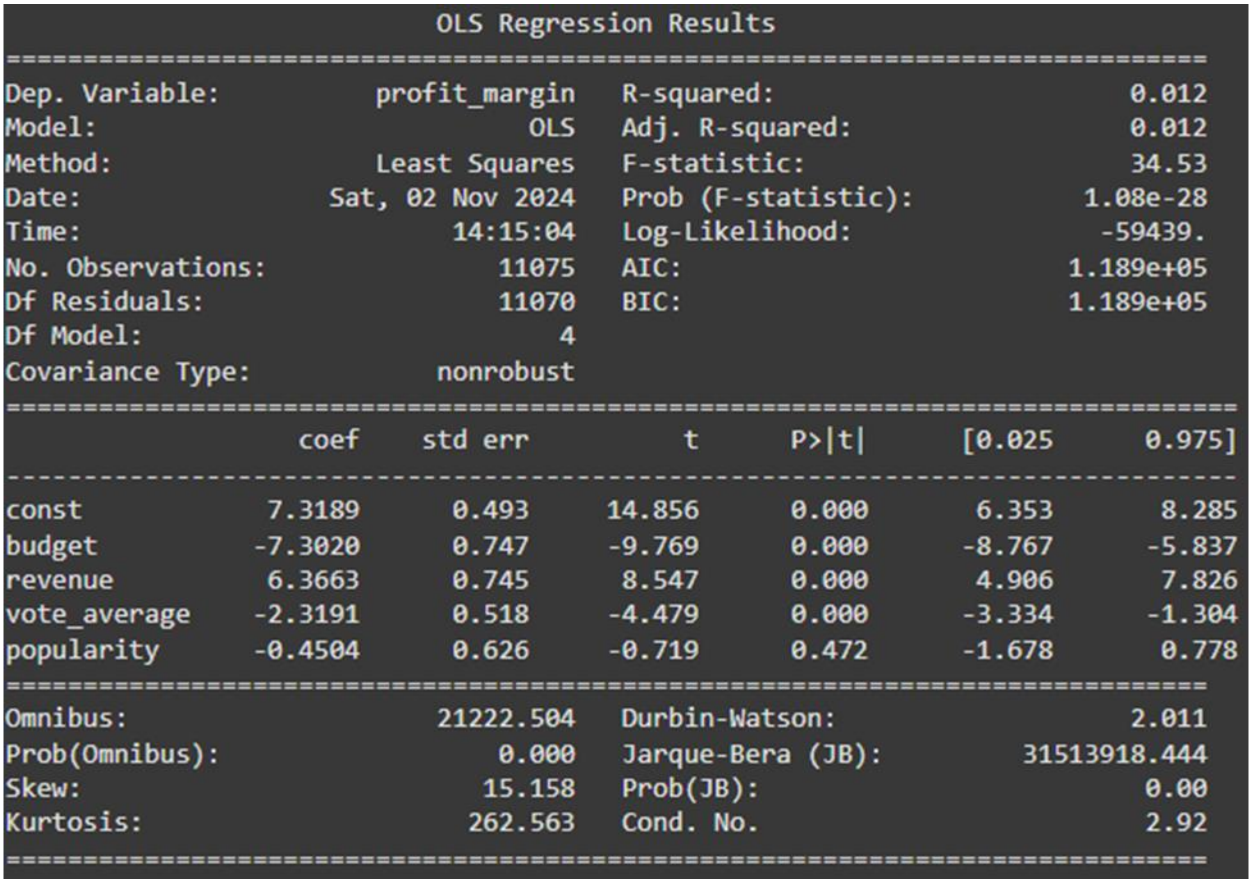
The dataset suggests that the two highest revenue earning movies are The Bee movie and TIkTokRizz party. Clearly this wrong. The highest revenue earning film of all time is Avatar, so this is just bad data.

In the top 10 movies with the highest budgets, the number 1 highest budget film is Lost in the Stars, which clearly is wrong. There's no way that movie had a budget that cost $400+ million unless they were using a different currency than what the other movies used. This means that we have a data issue because the data doesn't specify what currency the movies are in. To me this is the biggest concern. If we aren’t using the same currency for our films, then every film that isn’t using USD currency is bad data. This could make up a significant portion of our data. So its pretty clear to say that regardless what results we get from our predictive model, the model won’t be very trustworthy in predicting profit margin in the real world.

Lets see how well we can make this data predict the profit margin of films. Let’s perform an Ordinary Least Squares regression on the data. This is a linear model, it creates a line of best fit through the data by minimizing the sum of the squared differences between the observed values (actual data points) and the values predicted by the linear model.

We scaled the data, transform it and then fit it to the linear regression model.

We can look at OLS regression results to see how well our OLS regression model predicts our film’s profit margin.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          profit_margin   R-squared:                       0.012
Model:                            OLS   Adj. R-squared:                  0.012
Method:                 Least Squares   F-statistic:                     34.53
Date:                Sat, 02 Nov 2024   Prob (F-statistic):           1.08e-28
Time:                        14:15:04   Log-Likelihood:                -59439.
No. Observations:               11075   AIC:                         1.189e+05
Df Residuals:                   11070   BIC:                         1.189e+05
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            7.3189      0.493     14.856      0.000       6.353       8.285
budget          -7.3020      0.747     -9.769      0.000      -8.767      -5.837
revenue          6.3663      0.745      8.547      0.000       4.906       7.826
vote_average    -2.3191      0.518     -4.479      0.000      -3.334      -1.304
popularity      -0.4504      0.626     -0.719      0.472      -1.678       0.778
==============================================================================
Omnibus:                    21222.504   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         31513918.444
Skew:                          15.158   Prob(JB):                         0.00
Kurtosis:                     262.563   Cond. No.                         2.92
==============================================================================
```

Interpreting the results:

This is as low of an R-squared (and equally low adjusted R-squared) value as you can get. This means the model does a very poor job of explaining the variance in the profit margin. So there is weak explanatory power of the model regarding the profit margin.

The Prob F statistic very low at a statistically significant level, indicating that at least one of the predictor variables is a significant factor in predicting the profit margin.

The budget, revenue, and vote_average all have p-values that are statistically significant, so they have an effect on the profit margin.

A one standard deviation increase in revenue is associated with an increase of approximately 6.3663 units in the profit margin. So higher revenue leads to an increase in profit margin. Profit is derived from revenue, so this is not particularly useful result.

For each 1 standard deviation increase in the scaled budget, the profit margin is expected to decrease by approximately -7.3020 units. So its a negative relationship, higher budgets lead to lower profit margins. This implies that if you wanted to increase you film’s profit margin, its not as simple as just dumping more money into the budget.

A one unit increase in vote_average is associated with a decrease of approximately -2.3191 units in profit margin. This indicates that, counterintuitively, as a movie becomes more more highly rated, the profit margin might decrease. This is a very interesting finding. Following the logic to its extreme implications, the model predicts that you should make your film have a worse vote_average if you want your film to have a high profit margin.

It appears that popularity has a statistically insignificant effect upon the profit margin of the film. It appears that even if a film is popular on IMDB, it doesn't make a difference to the resulting profit margin of the film.

What does revenue and budget coefficient mean in common sense English? After performing a few calculations, we can see that. For every increase in $1 million in revenue, profit margin is predicted by the emodel to increase by 4%. For every increase in $1 million in budget, profit margin is predicted by the model to decrease by 20%.

Analysis of the audience voting activity.

The main purpose of the numeric analysis of the audience voting activity is to understand what motivates the viewers to vote for a movie and what factors influence the average voting scores. The viewer's voting action can be viewed at a high level as an expression of viewer's engagement, and willingness to share their opinion about the movie.

The dataset sourced from TMDB consisting of 1,127,777 movies will be used for this analysis. Some of the records in the source database are missing important metrics. Therefore, there is the need to use data filtering and cleaning technique to exclude the records that are not suitable for further analysis. In particular, all records with 0 and null values in the vote_count and vote_average columns were filtered out. The resulting subset includes 349,440 movies. Even though the resulting subset is significantly smaller than the initial dataset, 349k of movies is still sufficient volume to reliably identify trends and correlations in the data.

The analysis of voting activity will attempt to answer the following questions:

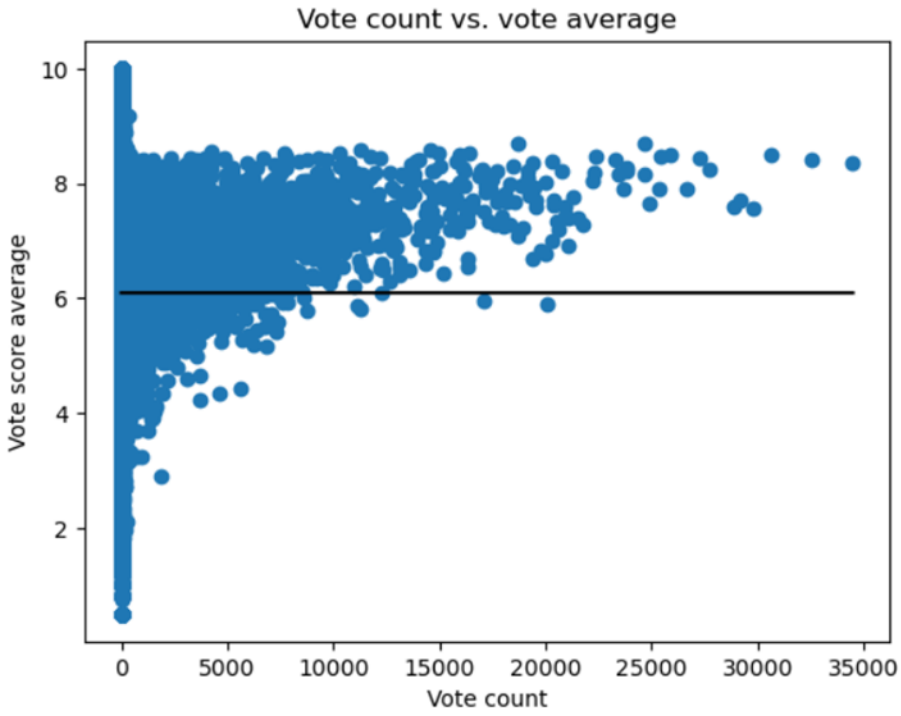
1. Is there a correlation between vote counts and the average scores?
2. Does the movie budget affect its voting count and average score?
3. Does the movie runtime have an impact on its voting count and average score?
4. What movie genres are most voted? Is it true that the most voted genres are also most highly ranked?
5. Does the movie score count and ranking depend on the production country?

The analysis will be performed and visualized using Python libraries pandas, numpy, statsmodels.api and matplotlib.

Is there correlation between vote count and average score?

Correlation between vote_count and vote_average shows if movies with high audience activity (vote_count) are more or less likely to have higher vote average.

We will use *statsmodels* Python library to perform the linear regression analysis. Selecting vote_count and vote_average and dropping all records with 0 values will be sufficient to prepare data for this analysis.



The diagram above shows a scatterplot and a linear regression line for Vote Count vs. Vote Average correlation.

The scatterplot clearly shows that the higher the vote count for particular movie is, the lower is the chance that this movie will receive low average score. None of the movies with vote count exceeding 10k received average score lower than 5.6.

Low rankings (below 4) can only be received by those movies that have medium and low vote counts (1k and less).

The linear regression of average vote scores from vote counts is roughly parallel to X axis and is positioned at the level y=6.09. Looking at the scatterplot diagram, we would expect that the linear regression line would be trending more upward. However, this is not the case because the density of the points in the left sector of the diagram is significantly higher and has the determining influence on the incline angle of the regression line. Points located in the middle and top right sector of the scatterplot are very scarce and they have lower influence on the regression line.

Therefore, we can conclude that there is a negligible regression of average vote scores from vote counts.

Global vote average

Global Vote Average will be calculated as weighted average of the movie vote scores scaled by vote count per each movie.

The result tells us that when taking into account all the votes given to all the movies in the dataset, the average score given by one vote is **~6.83**. The practical meaning of this metric is that viewers on average tend to give high scores (rather than low scores) when they vote for the movies.

Does the audience engagement depend on the movie budget?

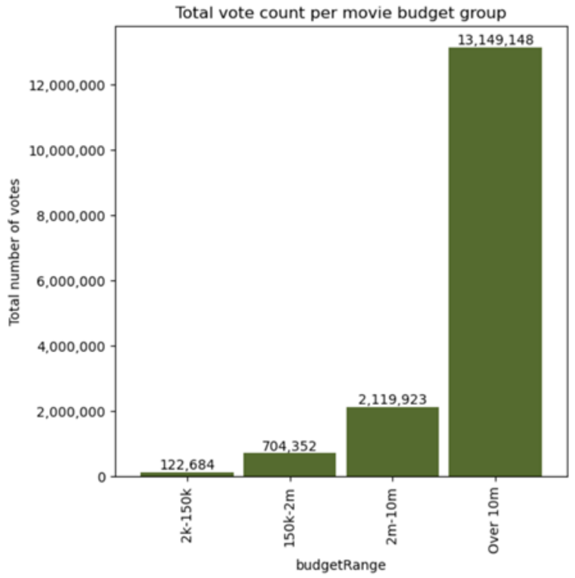
To answer this question we will group the movies according to their budgets and for each group we will identify the average vote count per movie.

For this analysis, we will need to take into consideration the column *budget* that has a lot of values requiring the cleaning. For example, there are budgets exceeding 500m and budgets below 100, that are not realistic values. To make sure the analysis is performed on the accurate data, we will set maximum (450,000,000) and minimum (2,000) budget thresholds and remove all the records that do not fit into this range.

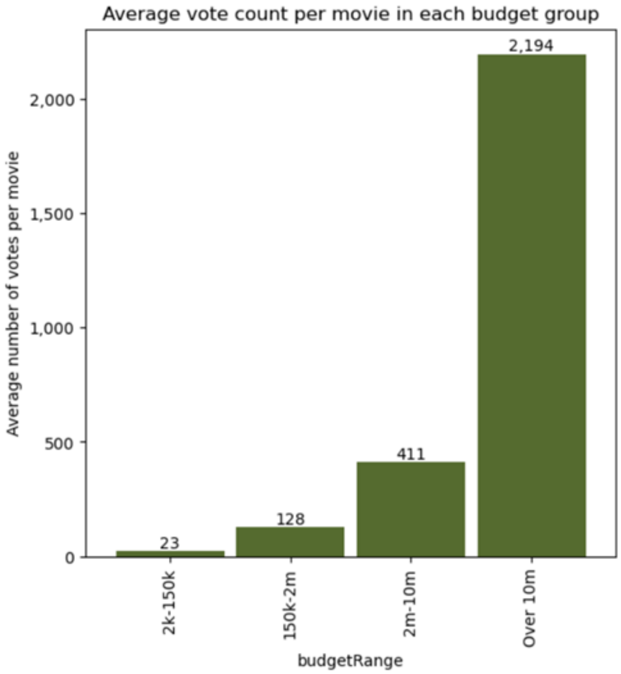
The size of dataset after cleanup is pretty low at 22,061 records. In this case, it is preferable to have small dataset of reliable data rather than large set with questionable budget values.

The distribution of the budget amount values in the resulting subset is not even. Therefore, we will use the *pd.qcut* function for defining 4 budget amount range groups to make sure that each group contains equal amount of movies in it.

By applying the qcut function, all the movies were split in 4 equal sized groups depending on their budget amount. The size of one group is around 5,500 movies.



The histogram 'Total vote count per movie budget group' visualizes total count of votes given to movies in each budget group. Majority of votes are being received by the movies with budgets over 10m (13.1m votes). The second largest total vote count is given to movies with budget between 2m and 10m (2.1m votes). Movies with budgets below 2m get significantly lower vote count when compared to the higher budget groups.



The histogram 'Average vote count per movie in each budget group' presents the average vote count per movie for each budget range group. The results are consistent with the previous diagram.

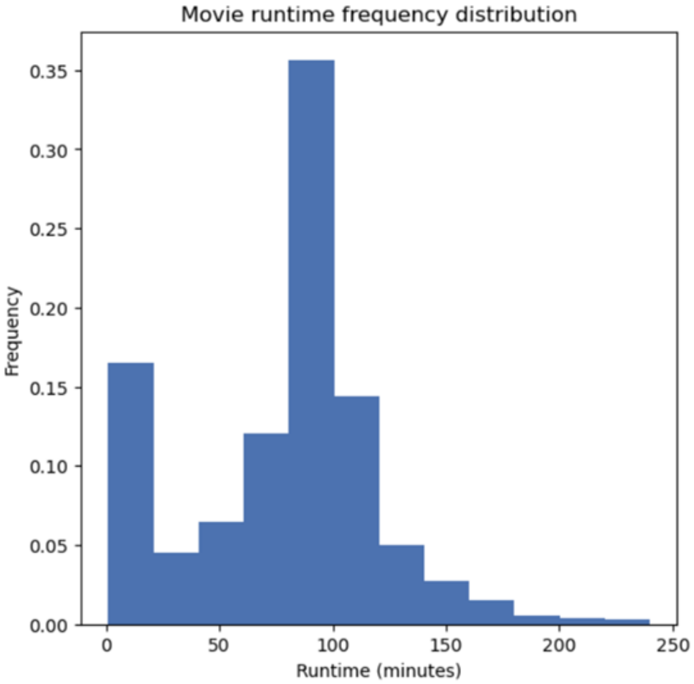
- The highest vote count is given to movies with budget over 10m - 2,194 votes per movie
- The lowest vote count is given to movies in the lowest budget group (2k-150k) - 23 votes per movie.

To summarize, movies with higher budgets attract more votes from the audience. A high budget movie receives 95x votes compared to a movie in a low budget group.

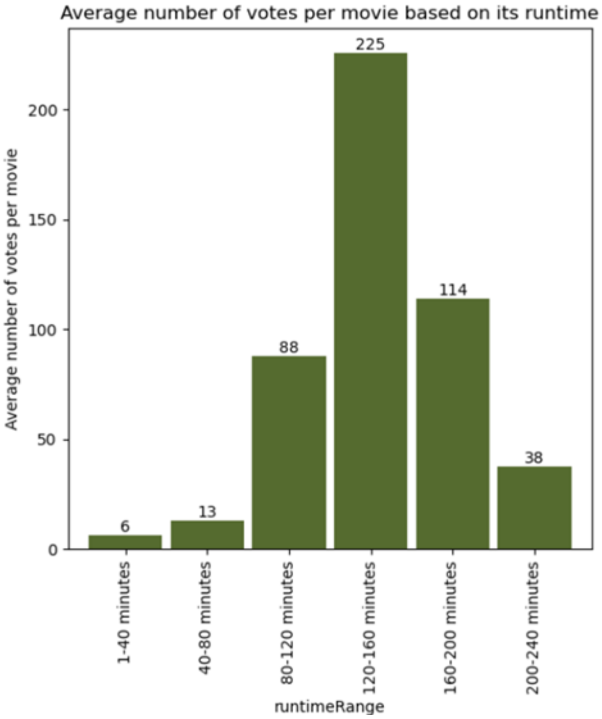
Does the audience engagement depend on the movie runtime?

The purpose of this analysis is to determine if vote counts and scores depend on the movie runtime.

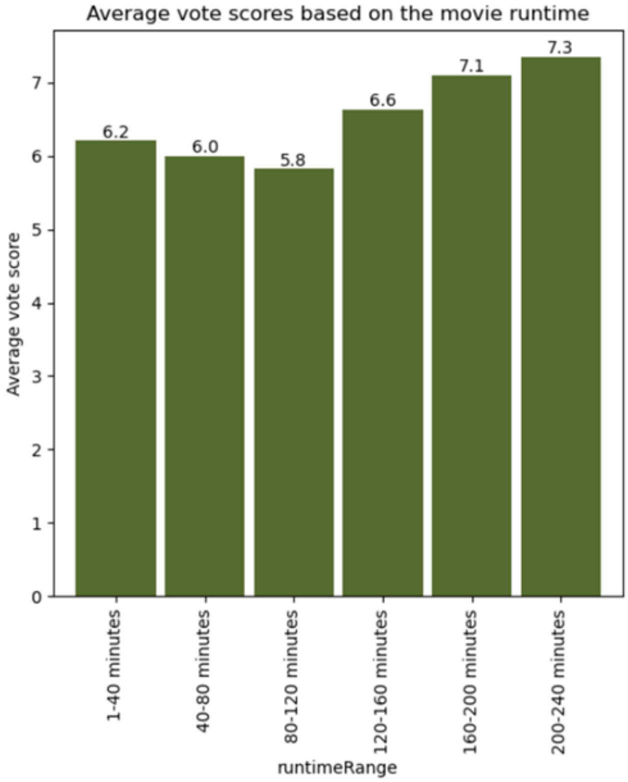
This analysis will take into account the *runtime* column from the source database. During the clean-up process we will remove all records where runtime value is equal or less than 0 and above 240 minutes. Vast majority of movies will fall into this interval. There was small share of movies in the dataset with runtime over 240 minutes. However, these records represent mini TV series, and they are not classified as standard movies, and therefore are excluded from the scope of this analysis.



To better understand the data, we first visualize the frequency distribution of the movie runtimes. As we can see from the diagram "Movie runtime frequency distribution", the most frequently produced movies are those with runtimes between 80 and 100 minutes (35 per cent frequency rate). The least frequently produced are movies with runtime over 180 minutes (around 1 per cent frequency rate). The interesting observation is movies with runtime between 0 and 20 minutes are fairly common (16 per cent frequency rate).



All the movies from the source dataset were split into 6 groups depending on their runtime. The histogram 'Average number of votes per movie based on its runtime' helps to discover the movies that receive lowest and highest vote counts based on their runtime. Lower vote counts are usually given to movies with runtimes between 1 and 80 minutes. The highest vote counts are given to movies with runtimes between 120 and 160 minutes. Average vote counts are given to movies with runtimes 80-120 and 160-240 minutes.



When looking at the diagram 'Average vote scores based on the movie runtime' we can also see some differences in vote scores between movies in different runtime groups. Movies with runtime 200-240 minutes are most highly rated (7.3) while movies with runtimes 80-120 minutes are scored the lowest (5.8).

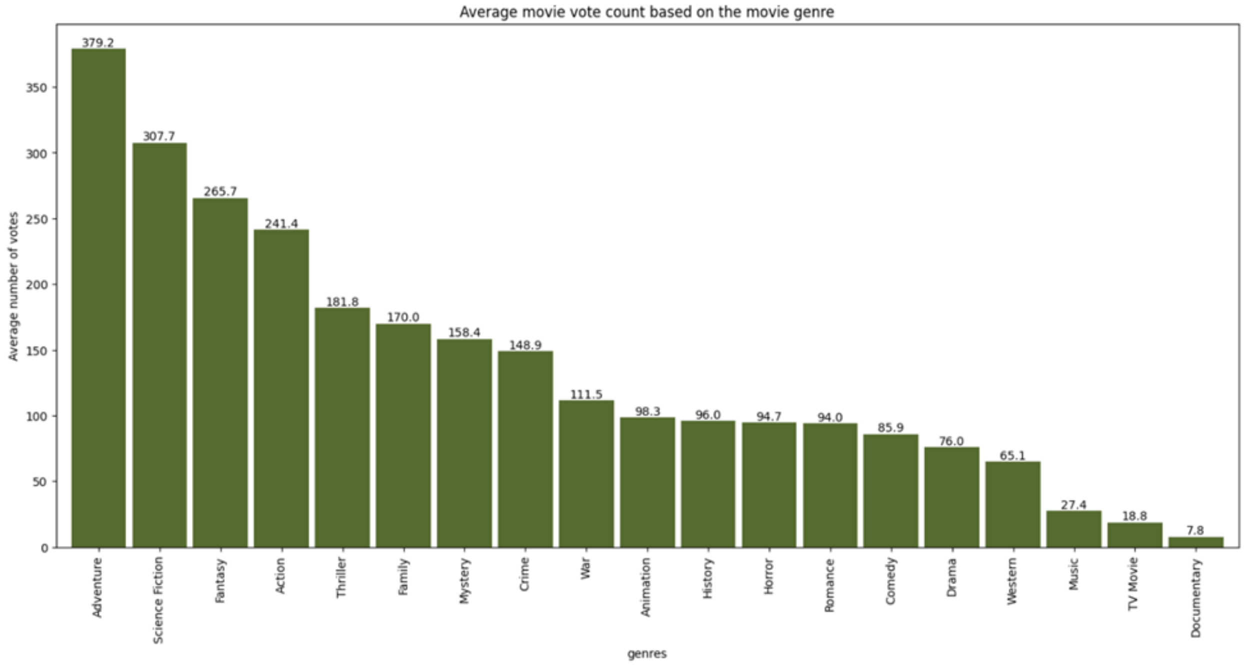
To conclude the movie runtime analysis, there is dependency of vote counts and scores on the movie runtime. Movies with runtimes between 120 and 160 minutes are most frequently voted, and movies with runtimes between 200 and 240 minutes are most highly scored.

What movie genres are most voted?

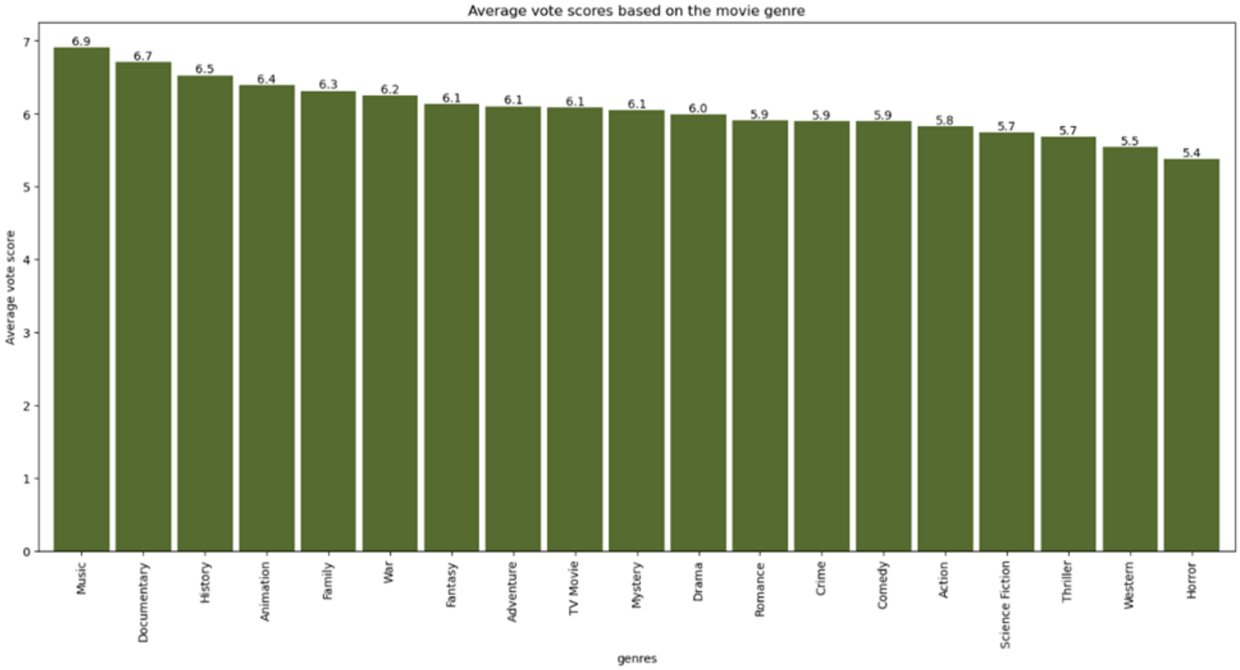
In this section we will identify the relationship between movie genre and its vote scores and vote counts.

There are 19 movie genres. One movie can belong to several movie genres at the same time, in which case value in the column 'genre' will contain a CSV list with applicable genre names (for example, 'Action, Science Fiction, Adventure'). In order to perform the analysis, we will decompose the records that have multiple genre names in the 'genre' column. One record with multiple genre names will be transformed into multiple records where each record will have only one genre name. For example, if a movie with id 123 belongs to genre 'Action, Adventure, Science Fiction', then as a result of transformation we will have 3 records: 'id 123 genre Action', 'id 123 genre Adventure', 'id

123 genre Science Fiction'. Normalization process increases number of records in dataset from 290k to 504k. All records will be grouped by single genre name and vote counts and vote scores will be calculated on each group for further comparison.



Based on the resulting diagram, there are definitive leaders that receive the highest average vote counts. Adventure, Science fiction, Fantasy and Action movies on average receive more than 200 votes per movie. The lowest vote counts (below 28 votes per movie) are received by movies in Music, TV Movie and Documentary genres.



The highest average vote scores are given to movies in Music, Documentary, History and Animation genres. The lowest scores (below 5.7) are assigned to Science Fiction, Thriller, Western and Horror movies.

Some movie genres express negative correlation between vote counts and vote scores. For example, Science Fiction movies are among the most highly voted (308 votes per movie on average), and at the same time they are among the least scored movies (average score 5.7). Music and Documentary movies express similar type of negative correlation. They are the most highly scored (6.9 and 6.7 respectively) and they are the least voted at the same time (27 and 8 votes per movie respectively).

Relation between movie production country and vote counts/scores

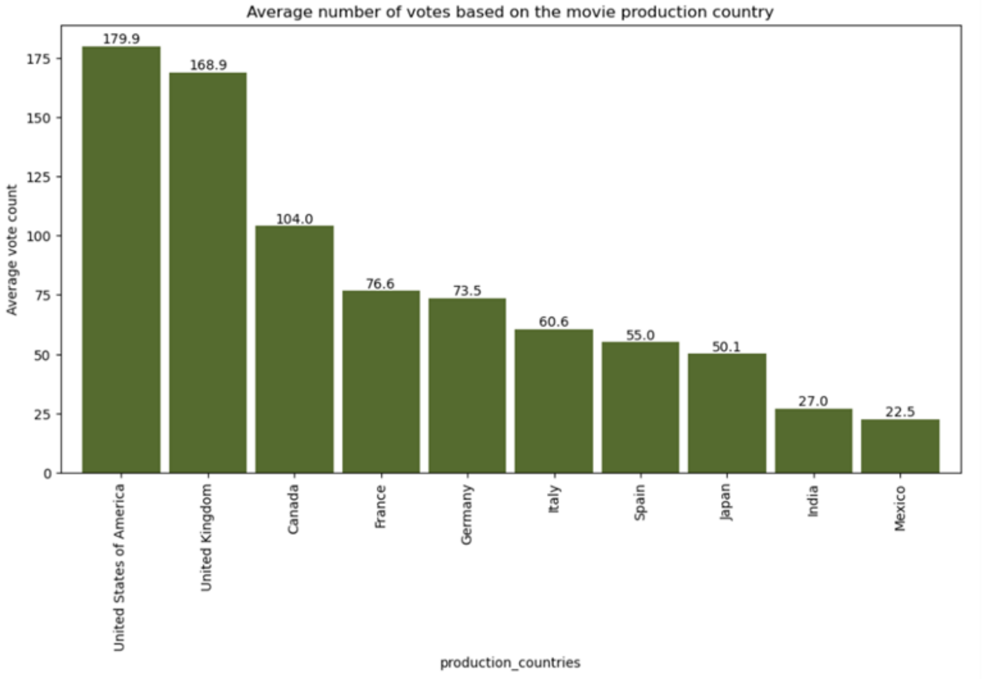
Voting activity can be influenced by such factor as the producing country. Some countries have advanced technologies and experience in movie production, and therefore movies made in these countries are better positioned to receive higher votes and scores.

There are many countries in the world that produce movies. Some countries produce negligible number of movies, so taking them into account doesn't make sense. For the vote count and average score by country analysis we will calculate the number of produced movies by each country and will identify and analyze top 10 movie producing countries.

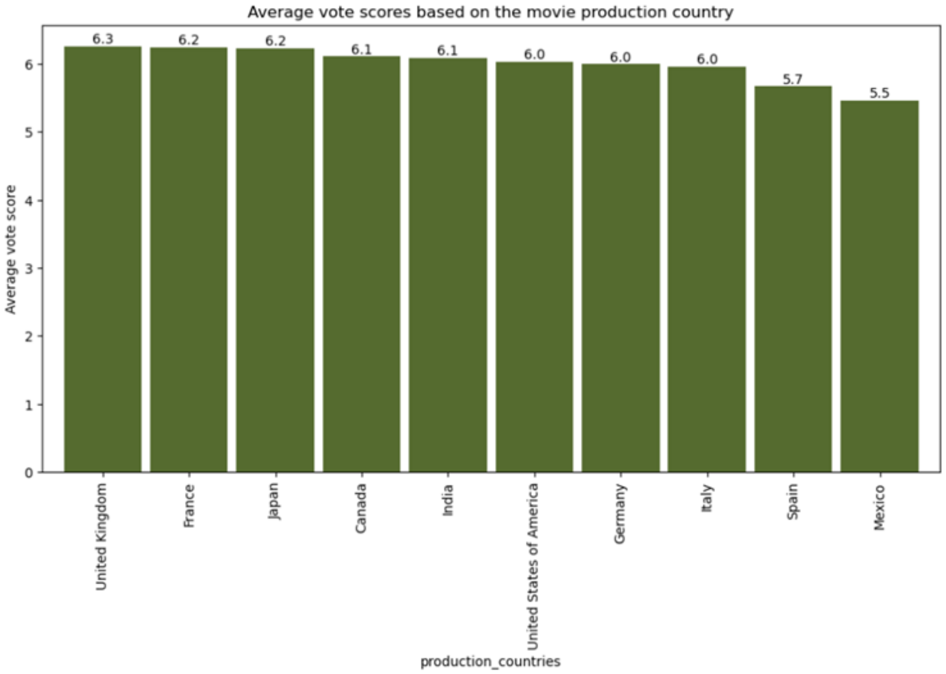
One movie can be produced by several countries. We will normalize the corresponding records and decompose them so that each resulting record corresponds to a single producing country.

The top 10 selected countries for analysis (based on the volume of movie production) are:

- United States of America (94,499)
- France (24,499)
- United Kingdom (19,022)
- Germany (16,370)
- Japan (14,004)
- Italy (11,131)
- Canada (11,099)
- India (10,008)
- Spain (8,504)
- Mexico (8,304)



There is significant difference in a number of votes for movies depending on the country of production. The most voted movies are produced in the United States of America and in the United Kingdom (180 and 169 votes per movie respectively). The least voted movies are coming from India and Mexico (27 and 25 votes per movie respectively).



While voting score depends on the country of production to lower extent, there are still some differences. The most highly scored movies are produced in the United Kingdom (6.3) followed by France and Japan (6.2) while the lowest scored movies are produced in Mexico (5.5).

Overall, United Kingdom, United States of America, Canada and France take leading positions in both, the vote counts and average scores for movies produced in these countries.

Analysis - Historical trends of budget, revenue and popularity of genres

Looking at trends and profitability of the film industry from inception to its current state using the TMDB dataset v11

Study:

- Understand the growth of the film industry and how popularity of different genres have influenced investment and revenue

Data Cleansing: Initial set of 1.1M rows reduced to 9**K useful movies**!

- Removed movies with no titles
- Removed movies release dates or outside 1915-2025 from Birth of a Nation to latest released movie
- Removed movies with $0 or NAN budgets and revenue as this is a financial analysis
- Removed movies with Votes less than 100 as this were mostly junk movies
- Removed Adult movies
- Removed several cherry-picked off-skew movies

Data Transformation:

- Since this is a study over time, categorized the data into decades based on release date from 1910 to 2020
- As this is a study over time looking at financial trends, corrected all historical $$ to 2024 CPI to account for inflation. It was highly problematic to connect to CPI data at runtime

so was able to extrapolate the historical data to match the periods analyzed locally. This was applied to the budget and revenue creating new data elements adjusted_budget and adjusted_revenue that will be used for all the analysis.

  - For clarity, there are two known problems with this approach
    - The data is corrected at the decade level. By example, a value from 1951 will get the same treatment as a value from 1958
    - While budget is static at a point in time, revenue is cumulative over time. However, for simplicity, the same CPI factor was applied to revenue which led to interesting analysis :-)

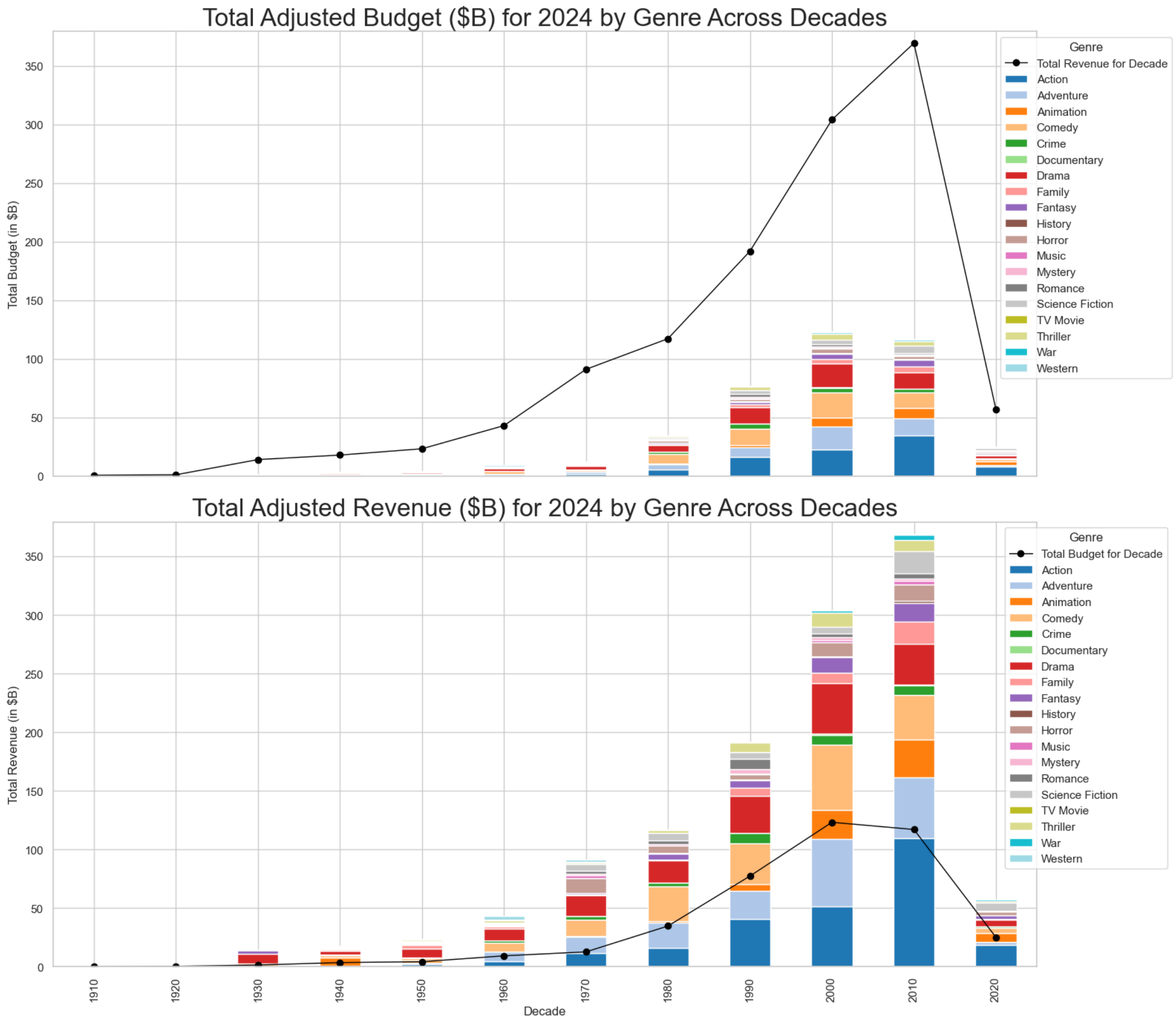
- The Genre which will be another key metric came as a collection with the dataset with some movies identifying as many as 19 unique genres. While this is accurate from a move keyword and indexing perspective, it caused havoc to genre based analysis. The Genre collections were broken down into individual columns, each identifying a unique genre. In the end, on the first, which consistently showed to the primary genre, was used.

Budget vs Revenue over Time – Macro View:

The first analysis was breaking the movies down by decades and plotting them by decades but divided by Genre (Genre_0).

Here, we used a multi-facetted stacked bar graph with a bar for each decade. There are two subplots, the first showing budget broken down by decade and genre (cummulative) with a line index showing the total revenue (all genres) for that decade. The second reverses the view and shows the revenue generated by each genre across decades with a line illustrating the total budget for that decade.

All revenue and budgets use adjusted values for 2024

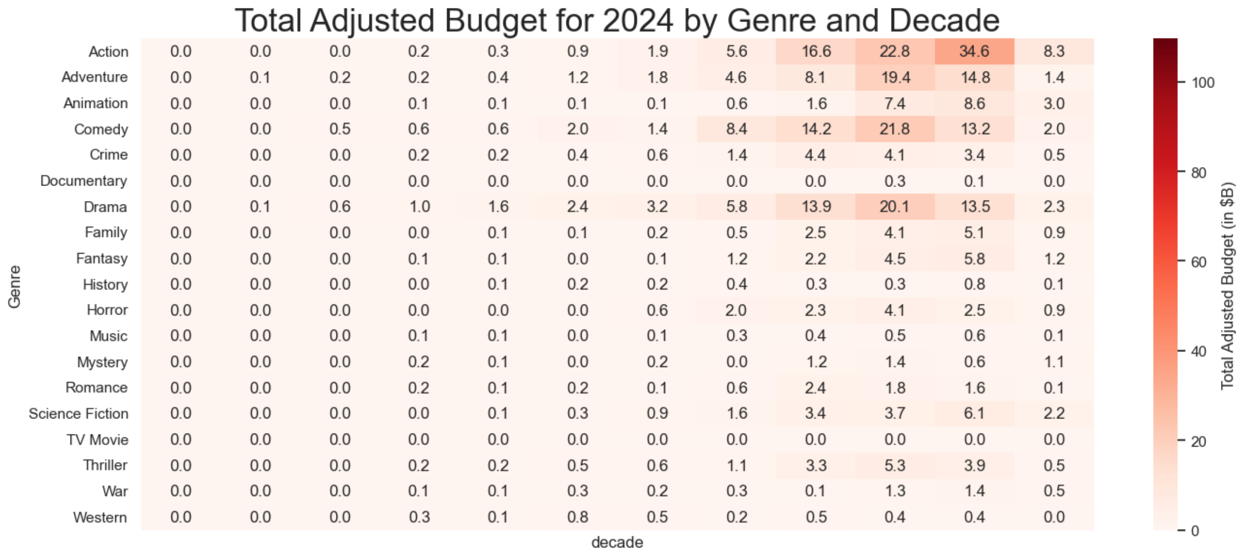


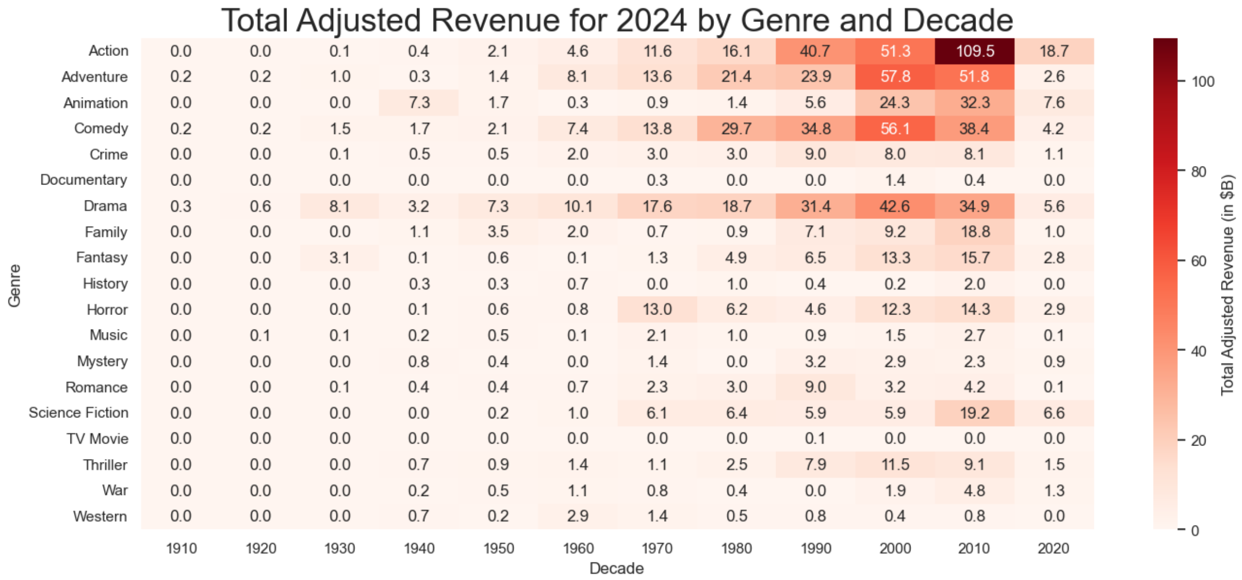
This shows a very clear indication that the movie industry as a whole has been dramatically increasing since inception. BUT there is a dramatic fall for 2020s for 2 reasons:

- We only have 4yrs of data for 2020s
- Even with 4yrs of data, the budgets and revenue are drastically lower than 40% of 2010s. Keeping at the same ration of increment enjoyed by the movie industry since inception, the 2024 numbers (almost 5yrs) should be at approximately $218B in revenue as opposed to the current $55B. We see this trend diminishing since the beginning of 2020 right alongside the COVID lockdowns

At a macro level, this model also shows key “bread and butter” genres that a sure hits for the movie industry … Action, Adventure, Comedy, Drama. The investments in these genres are high and the revenues coming in are very profitable.

Budget vs Revenue over Time – Drilling Down:





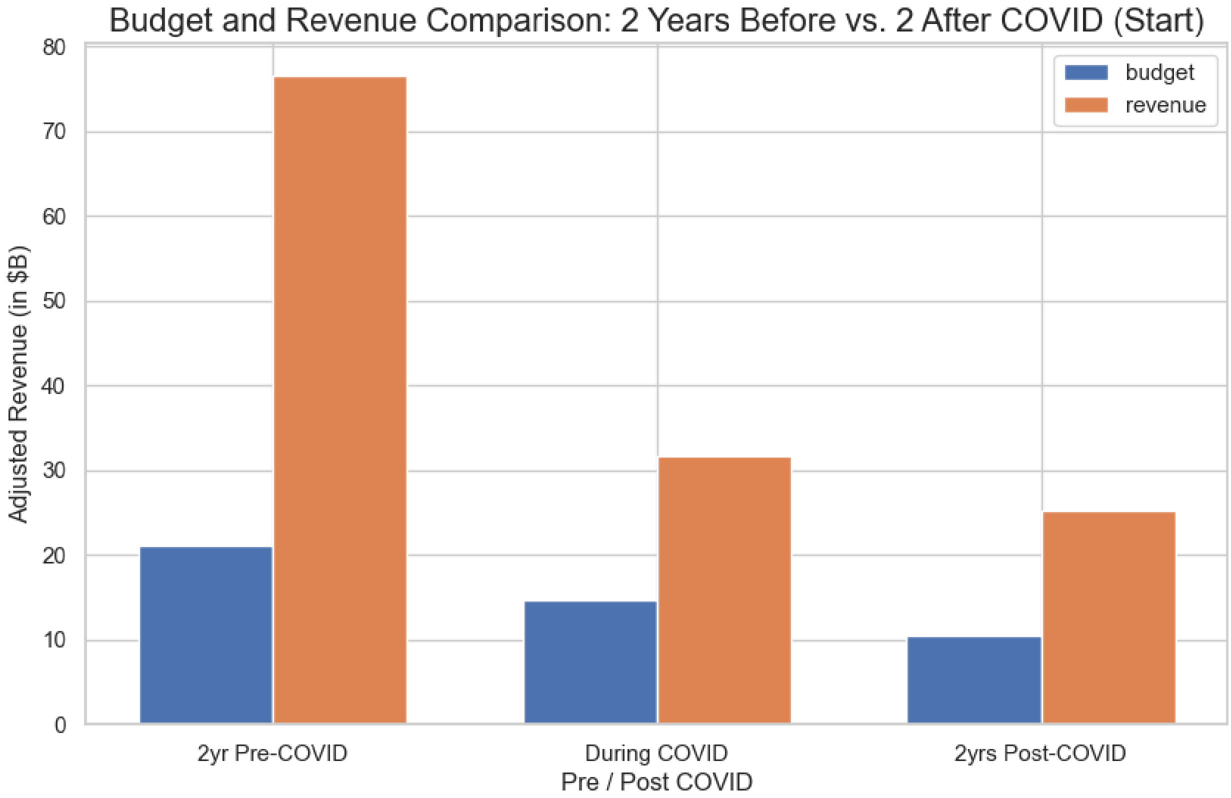
This success if further validated as we start diving into the data that shows continuous growth success in Genres Action, Adventure, Animation, Comedy, Drama, Documentary, Horror, Science Fiction since the 1970s. In reflection, these genres also see the largest budgets.

Also, very clear is there is an **ABNORMAL** massive downturn in both budget and revenue in this decade. With five years of data between 2020 and 2024, previous trends would project that by now, generation of approximately $210B would have been generated … however, only $60B revenue has been generated. This is also apparent in the budgets to date. To date, projection from previous decades would show an investment of $60B while the actual investment to date is $30B.

These numbers also shows that the movie industry is TRYING to invest. The budgets are just at half their previous decades spend, however the revenue is just above a quarter of previous decades.

Budget vs Revenue over Time – COVID:

In order to drill into this, the budget/revenue information of 2 yrs before COVID, during COVID and 2 yrs post COVID lockdown was aggregated.

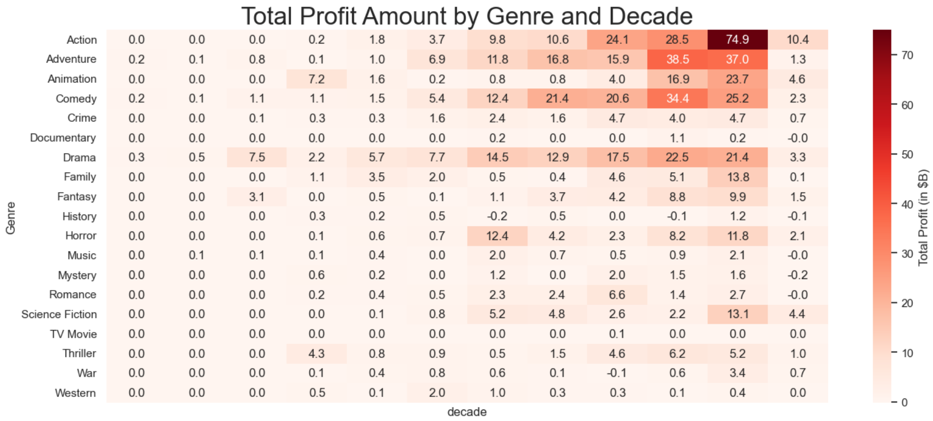


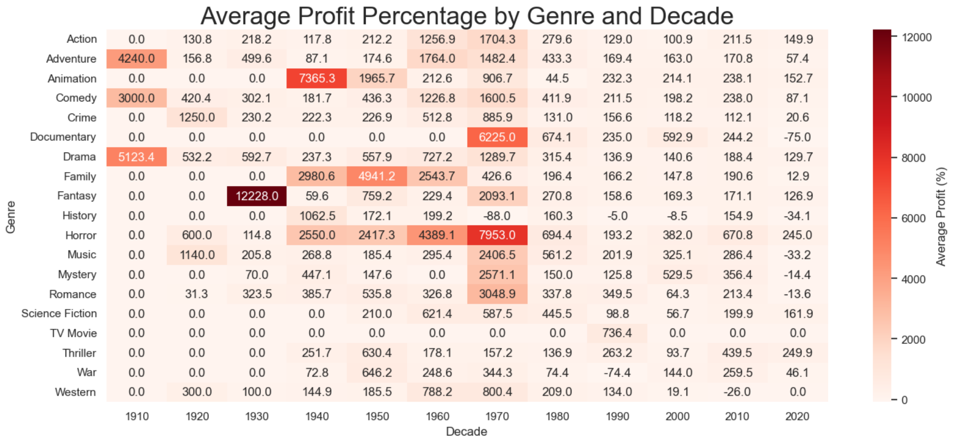
The number are more devastating than initially seen in the macro decade view! The post COVID numbers (2yrs of data) are barely projected to catch up to the DURING COVID (2.5yrs of data) numbers!!

We stopped going to theatres during COVID … this is a known fact. But we never went back after COVID!

Looking at Profit from Different Lenses:

Coming back to revenue over time and its consistent increase, we dissect the data to get a stronger understanding on how the revenue increase tracks over time.

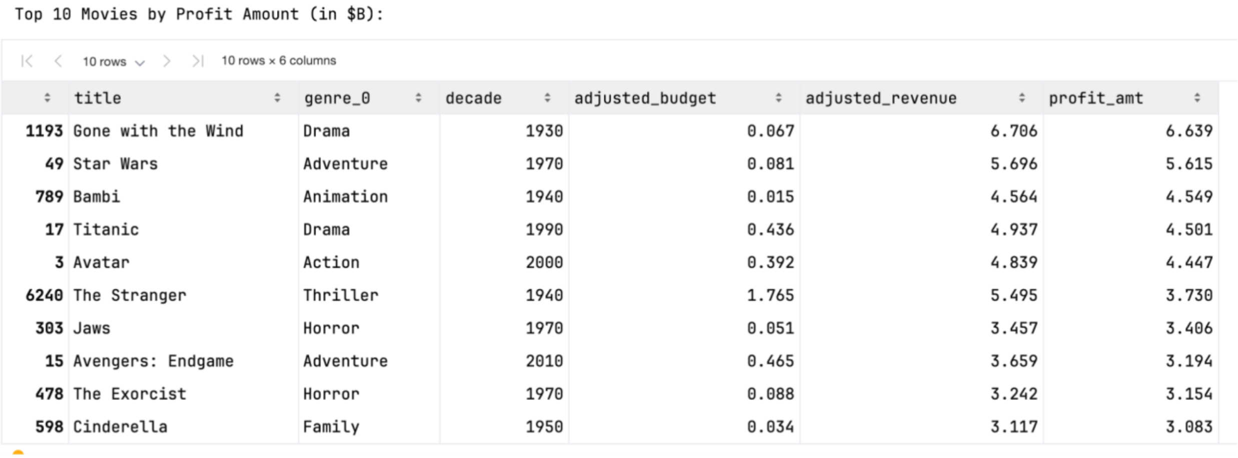




In evaluating profit based on value returned on initial budget, we are sure to think that the big budget movies of the 21st century top the bill. If you look at the revenue amounts, they do. However, if you take a different perspective and look at the % of return based on initial budget, we see a very different story.

The big budget movies can return significant returns, but the big winner in percentage returns historically have landed on smaller budget movies in the 1940s-1970s.

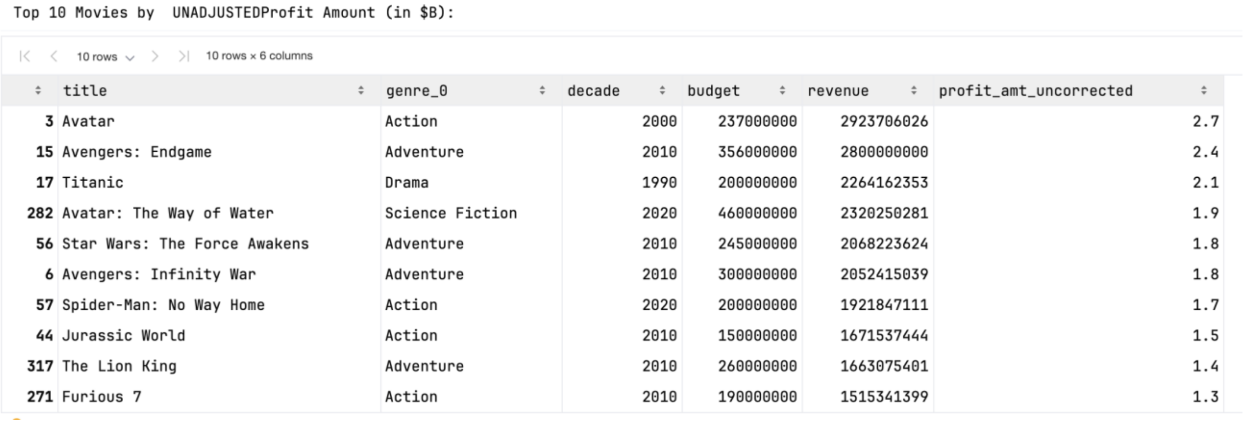
Looking into these movies in more detail as the highest grossing and the highest percentage, we get these results.



Top 10 Movies by Profit Amount (in $B):

	title	genre_0	decade	adjusted_budget	adjusted_revenue	profit_amt
1193	Gone with the Wind	Drama	1930	0.067	6.706	6.639
49	Star Wars	Adventure	1970	0.081	5.696	5.615
789	Bambi	Animation	1940	0.015	4.564	4.549
17	Titanic	Drama	1990	0.436	4.937	4.501
3	Avatar	Action	2000	0.392	4.839	4.447
6240	The Stranger	Thriller	1940	1.765	5.495	3.730
303	Jaws	Horror	1970	0.051	3.457	3.406
15	Avengers: Endgame	Adventure	2010	0.465	3.659	3.194
478	The Exorcist	Horror	1970	0.088	3.242	3.154
598	Cinderella	Family	1950	0.034	3.117	3.083

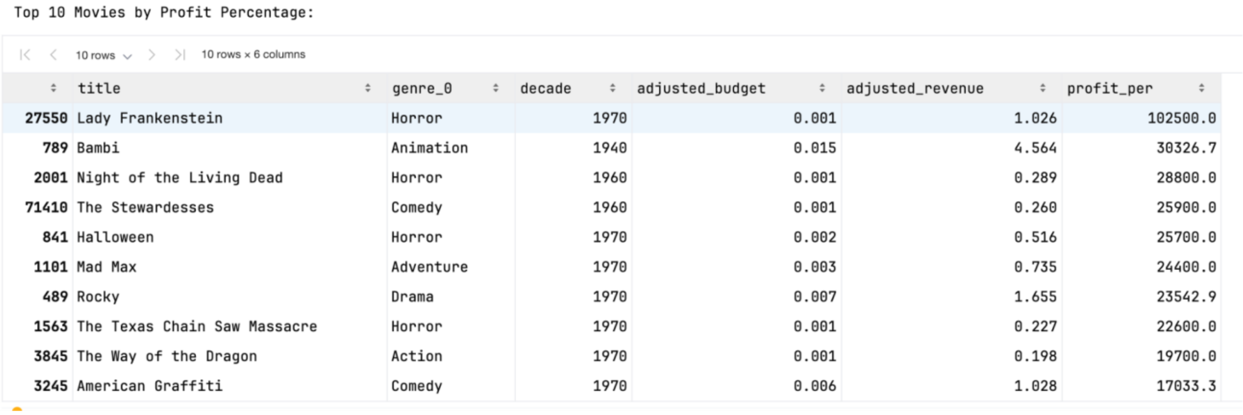
Above, we see the highest revenue movies when the budget and the revenue has been adjusted for time … Gone with the Wind is the movie with the highest grossing amount when adjusted for time



Top 10 Movies by UNADJUSTEDProfit Amount (in $B):

	title	genre_0	decade	budget	revenue	profit_amt_uncorrected
3	Avatar	Action	2000	237000000	2923706026	2.7
15	Avengers: Endgame	Adventure	2010	356000000	2800000000	2.4
17	Titanic	Drama	1990	200000000	2264162353	2.1
282	Avatar: The Way of Water	Science Fiction	2020	460000000	2320250281	1.9
56	Star Wars: The Force Awakens	Adventure	2010	245000000	2068223624	1.8
6	Avengers: Infinity War	Adventure	2010	300000000	2052415039	1.8
57	Spider-Man: No Way Home	Action	2020	200000000	1921847111	1.7
44	Jurassic World	Action	2010	150000000	1671537444	1.5
317	The Lion King	Adventure	2010	260000000	1663075401	1.4
271	Furious 7	Action	2010	190000000	1515341399	1.3

Above, we see the same matrix but removing the adjustment for time. In this case, Avatar is the highest ranking and ‘Gone with the Wind’ doesn’t make the top 10!

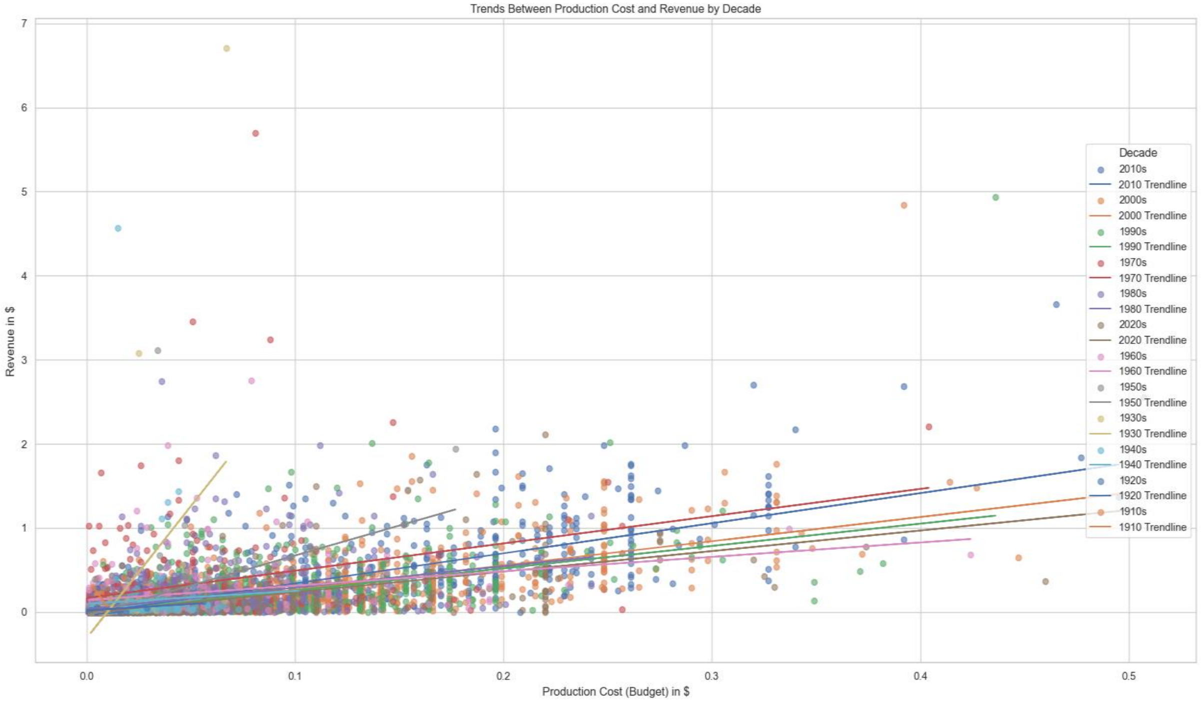


Top 10 Movies by Profit Percentage:

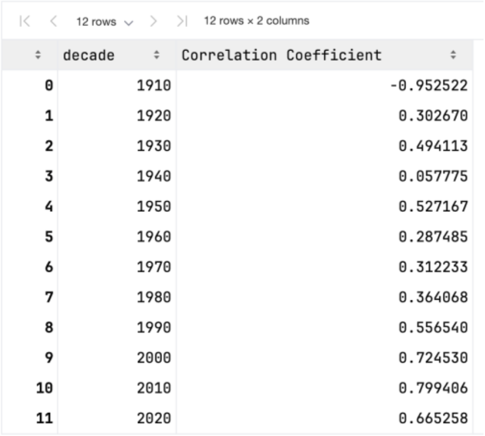
	title	genre_0	decade	adjusted_budget	adjusted_revenue	profit_per
27550	Lady Frankenstein	Horror	1970	0.001	1.026	102500.0
789	Bambi	Animation	1940	0.015	4.564	30326.7
2001	Night of the Living Dead	Horror	1960	0.001	0.289	28800.0
71410	The Stewardesses	Comedy	1960	0.001	0.260	25900.0
841	Halloween	Horror	1970	0.002	0.516	25700.0
1101	Mad Max	Adventure	1970	0.003	0.735	24400.0
489	Rocky	Drama	1970	0.007	1.655	23542.9
1563	The Texas Chain Saw Massacre	Horror	1970	0.001	0.227	22600.0
3845	The Way of the Dragon	Action	1970	0.001	0.198	19700.0
3245	American Graffiti	Comedy	1970	0.006	1.028	17033.3

Above, we look at it from a percentage profit perspective and the highest percentage profit goes to Lady Frankenstein and neither Gone with the Wind nor Avatar are in the Top 10! In fact, the highest profitable movies are the as inface VERY LOW BUDGET movies that happened get a cult following to drive their revenue!

Is there a historical correlation between Budget and Revenue?



History shows a weak to medium correlation between initial budget and revenue generated. However, the last 3 decades have shown a much stronger correlation … except 2020s!


12 rows
12 rows × 2 columns

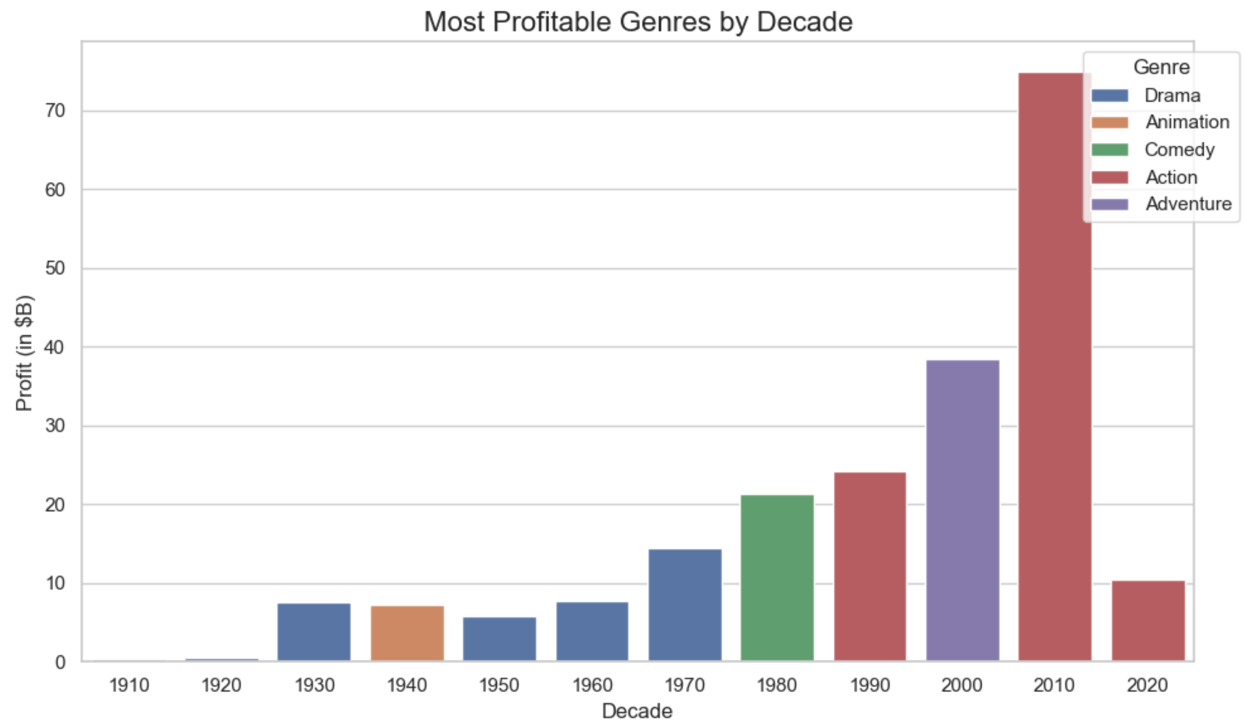
	decade	Correlation Coefficient
0	1910	-0.952522
1	1920	0.302670
2	1930	0.494113
3	1940	0.057775
4	1950	0.527167
5	1960	0.287485
6	1970	0.312233
7	1980	0.364068
8	1990	0.556540
9	2000	0.724530
10	2010	0.799406
11	2020	0.665258

Outside of an outlier in the 1950s, we can see the correlation between budge and revenue starts to increase in the 1970s starting from a medium correlation into the 2000s when we start seeing a strong correlation.

Again, the 2020s threw a wrench into the correlation, but there is still a medium strong correlation.

Summary

Through a historical view, we can see that budget has not always been strong predictor of revenue but there has been a lot better projection over the last several decades especially with favourable genres around the profitability of investments.



Language-Based Analysis of Movie Popularity

Report on Text Features, Sentiment Trends, and Insights

This report presents an analysis of the relationships between textual features (overview, tagline, keywords) and the popularity of movies. The dataset was analyzed to uncover patterns in word usage of marketing elements, sentiment, and thematic elements that differentiate popular, moderately popular, and less popular movies. The analysis aims to provide actionable insights into audience preferences based on linguistic and thematic trends.

1. Data Cleaning and Preprocessing

The dataset underwent several preprocessing steps to ensure the quality and relevance of the analysis:

1. **Removed Missing and Invalid Data:**
    a. Excluded movies without titles or release dates.
    b. Removed movies with zero or invalid vote counts.
2. **Filtered by Criteria:**
    a. Excluded adult movies.
    b. Focused on movies released before 2030.
3. **Text Standardization:**
    a. Removed punctuation and stopwords from text fields (overview and tagline).
4. **Derived Metrics:**
    a. Computed a *Weighted Vote Score* as a product of vote average and the logarithm of vote count.
    b. Categorized movies into three popularity levels:
        i. **Popular**: Top 20% by weighted vote score.
        ii. **Moderately Popular**: Middle 60%.
        iii. **Less Popular**: Bottom 20%.

The distribution of movies (with vote scores) across weighted vote categories indicates that the majority fall into the "Moderately Popular" category (191,857 movies), while significantly fewer are classified as "Less Popular" (66,686) or "Popular" (64,636).

2. **Word Frequency Analysis** of Movie Overviews and Taglines

2.1 Word Frequency in Movie Overviews

The overview text captures the essence of a movie's story and provides a key indicator of its themes and tone. The word frequency analysis reveals commonalities and distinctions between popular and less popular movies.

Top Words in Popular Movies' Overview:

1. **Dominant Themes:** Words such as "life" (13,839), "young" (8,387), "love" (7,193), "family" (7,160), and "friend" (6,778) dominate, reflecting a strong emphasis on relationships, personal journeys, and emotional resonance.
2. **Action and Agency:** Verbs like "find" (8,570), "get" (6,375), "take" (6,095), and "make" (3,923) highlight dynamic and engaging narratives.
3. **Diversity:** Words like "woman" (6,274) and "man" (6,979) indicate broad representation in storylines.

Top Words in Less Popular Movies' Overview:

1. **Thematic Overlap:** Similar words, including "life" (10,587), "young" (6,916), and "love" (6,757), suggest comparable foundational themes to popular movies.
2. **Lower Engagement:** The frequency of action-oriented words like "find" (5,722), "get" (5,458), and "make" (3,294) is notably lower, indicating potentially less dynamic narratives.
3. **Smaller Emotional Spectrum:** Family-focused words like "family" (4,672) and "friend" (5,134) appear less prominently.

Key Observations:

- Popular movies generally present a more dynamic and emotionally resonant narrative, as seen in the prevalence of action-oriented and relational keywords.
- Less popular movies lean towards similar thematic elements but with less intensity and variety.

2.2 Word Frequency in Movie Taglines

Taglines serve as a marketing hook, encapsulating a movie's appeal in a brief and engaging manner. The word frequency analysis provides insights into how popular and less popular movies differentiate in their promotional language.

Top Words in Popular Movies' Taglines:

1. **Emotional Connection:** Words like "love" (2,073), "story" (1,368), and "life" (1,304) emphasize relatability and emotional depth.
2. **Universal Appeal:** Words such as "world" (1,030), "time" (806), and "adventure" (495) suggest a broader, aspirational focus.
3. **Relational and Personal Hooks:** Words like "family" (506), "he" (496), and "woman" (612) indicate strong personal and character-driven stories.

Top Words in Less Popular Movies' Taglines:

1. **Reduced Engagement:** Lower frequencies of emotionally charged words like "love" (569) and "life" (351) reflect a less compelling narrative hook.
2. **Darker and Niche Focus:** Words like "death" (139) , "comedy" (131), and "murder" (122) suggest niche or darker themes, potentially appealing to smaller audiences.

Key Observations:

- Taglines for popular movies successfully combine emotional depth and universal themes, attracting a broader audience.
- Less popular movies often rely on niche or darker themes, which may limit their overall appeal.

2.3 Keyword Analysis

Keywords provide structured, metadata-like insights into movie themes and genres.

Top Keywords in Popular Movies:

1. **Diverse Representation:** Keywords like "woman director" (3,511), "lgbt" (931), and "gay theme" (910) point to inclusive and representative content.
2. **Storytelling Richness:** Keywords such as "based on novel or book" (2,868), "short film" (2,163), and "based on true story" (1,026) suggest depth and relatability.
3. **Universal Themes:** Terms like "love" (878), "friendship" (814), and "sports" (777) indicate widespread appeal.

Top Keywords in Less Popular Movies:

1. **Niche Interests:** Words like "softcore" (498), "pink film" (308), and "silent film" (345) suggest targeted, specific content.
2. **Experimental Genres:** Keywords such as "found footage" (207) and "stand-up comedy" (270) reflect experimental or less mainstream content.
3. **Darker and Localized Themes:** Terms like "world war ii" (159), "martial arts" (191), and "philippines" (211) suggest regional or specialized storytelling.

Key Observations: Popular movies emphasize inclusivity, universal themes, and engaging narratives, while less popular movies focus on niche genres and darker tones.

2.4 Insights and Implications

2.4.1 Commonalities Across Categories

Both popular and less popular movies share foundational themes of life, love, and relationships, indicating these are universal draws for audiences.

2.4.2 Distinctions in Popularity

Popular Movies: Highlight action, dynamic narratives, and a broad emotional spectrum. Taglines emphasize universal and aspirational themes, creating strong audience connections.

Less Popular Movies: Tend to focus on darker or niche themes, with less frequent use of dynamic or engaging keywords, potentially reducing broader appeal.

2.4.3 Implications for Stakeholders

For Writers and Producers: Incorporating universal themes and dynamic narratives while maintaining diversity can enhance movie appeal.

For Marketers: Using positive, inclusive, and adventure-focused language in taglines can optimize audience engagement.

2.5 Conclusion on Word Frequency Analysis

This expanded analysis confirms that linguistic choices in overviews and taglines influence a movie's popularity. While both popular and less popular movies share foundational themes of life, love, and relationships, more popular movies successfully combine dynamic and emotionally resonant storytelling with these universal themes, setting them apart from less popular counterparts. These findings can inform strategies for storytelling, content creation, and marketing to better align with audience preferences.

3. N-Gram Analysis of Movie Overviews and Taglines

N-gram analysis identifies frequently occurring word combinations in in movie overviews and taglines, providing deeper insights into thematic patterns and storytelling elements of popular versus less popular movies. This section analyzes bigrams (two-word phrases) and trigrams (three-word phrases), highlighting the linguistic choices in both storytelling and promotional materials.

3.1 N-grams in Movie Overviews: Diverging Narrative Styles

The language in movie overviews reveals how popular and less popular movies present their narratives.

Popular Movies:

- Popular movies heavily rely on dynamic and engaging phrases like *"fall love"* and *"high school"* that evoke emotions and energy. The recurring use of *"new york"* and *"world war"* suggests a preference for universally relatable settings and dramatic historical backdrops.
- Trigrams such as *"new york city"* and *"must find way"* reflect the appeal of stories rooted in transformation, perseverance, and urban settings that resonate broadly with audiences.

Less Popular Movies:

- Less popular movies, while also focusing on themes like romance (*"fall love"*) and youth (*"young man," "young woman"*), exhibit a narrower narrative scope. Phrases such as *"tell story"* and *"year ago"* suggest a less immersive or action-driven storytelling approach.
- Trigrams like *"mixed martial art"* and *"art event held"* hint at experimental or niche content that may limit mass appeal.

Key Contrast: Popular movies tend to center on universal themes of love, growth, and resilience, presented in dynamic, action-oriented phrasing. In contrast, less popular movies focus on more static or niche narratives, potentially reducing their emotional engagement and relatability.

3.2 N-grams in Movie Taglines: Marketing for Engagement

The tagline analysis highlights how popular and less popular movies market themselves to their audiences.

Popular Movies:

- Taglines for popular movies aim to inspire and evoke strong emotions. Bigrams like *"true story"* and *"love story"* emphasize authenticity and emotional depth, while phrases such as *"get ready"* and *"far would"* create excitement and a sense of adventure.
- Trigrams like *"dream come true"* and *"every family secret"* underline aspirational and dramatic hooks, effectively drawing in a wide audience by promising transformative or emotionally charged experiences.

Less Popular Movies:

- Taglines for less popular movies often reflect narrower or darker themes. While phrases like *"love story"* and *"true story"* are still present, their frequency is significantly lower. Words such as *"animation short"* and *"serial killer"* indicate a focus on niche genres or unconventional storytelling.
- Trigrams like *"beloved puppy find"* and *"thing better left"* suggest storytelling that appeals to specific or experimental tastes, which may not resonate broadly.

Key Contrast: Popular movie taglines successfully combine universal themes with aspirational and emotionally charged language, creating a broad and relatable appeal. Conversely, less popular movies rely on simpler or niche-specific phrasing, which may lack the universal resonance needed to attract a wide audience.

3.3 Insights

3.3.1 Thematic Breadth vs. Niche Appeal

Popular movies leverage broad and universally appealing themes such as love, resilience, and transformation. Their frequent use of dramatic or action-oriented phrases ensures widespread relatability. Less popular movies, while occasionally overlapping in thematic focus, lean towards niche or experimental narratives, which may limit their audience to specific interests or preferences.

3.3.2 Emotional Resonance

Taglines in popular movies prioritize emotional hooks and promises of excitement or personal transformation, ensuring immediate audience connection. Less popular movies often forgo these hooks in favor of specialized or unconventional themes, which may fail to captivate a broader demographic.

3.3.3 Strategic Messaging

Popular movies display a clear advantage in crafting both their narratives and promotional language to maximize engagement. By emphasizing dynamic, relatable, and aspirational phrasing, they appeal to a wide audience. Less popular movies often miss this opportunity, focusing instead on content that may only resonate with smaller, targeted groups.

3.4 Conclusion

The n-gram analysis underscores significant differences in how popular and less popular movies position themselves through language. Popular movies use broad, dynamic, and emotionally engaging phrasing in both overviews and taglines, enhancing their universal appeal. Less popular movies, while sometimes thematically similar, tend to rely on static or niche language that limits their broader resonance. These findings emphasize the importance of strategic linguistic choices in storytelling and marketing to connect with diverse audiences effectively.

4. Sentiment Analysis: Emotional Tone in Movie Overviews and Taglines

Sentiment analysis examines the emotional tone of text, providing insights into the affective appeal of movie overviews and taglines. By comparing average sentiment scores across popularity categories—Popular, Moderately Popular, and Less Popular—this analysis explores how emotional tone correlates with audience appeal.

4.1 Sentiment Trends in Overviews

Movie overviews set the stage for audience expectations, often reflecting the emotional and narrative tone of a movie.

Popular Movies:

- **Overview Sentiment Score:** -0.0898 (lowest across categories)
- Popular movies exhibit a more negative average sentiment in their overviews compared to moderately popular (-0.0323) and less popular (-0.0418) movies.
- This trend suggests that popular movies frequently employ dramatic, intense, or emotionally charged language, aligning with themes of struggle, transformation, or adversity that resonate broadly with audiences.

Less Popular Movies:

- **Overview Sentiment Score:** -0.0418 (closer to neutral)
- Less popular movies lean toward less emotionally intense narratives, with a milder use of negative language. This could reflect simpler or more static storytelling that lacks the dramatic tension seen in popular movies.

Key Contrast: The stronger negative sentiment in popular movies likely reflects the use of high-stakes and dramatic themes that engage audiences more effectively. In contrast, less popular movies' milder sentiments may fail to evoke the same emotional investment.

4.2 Sentiment Trends in Taglines

Taglines serve as a movie's marketing hook, often leveraging positive sentiment to entice viewers.

Popular Movies:

- **Tagline Sentiment Score:** 0.0412 (highest across categories)
- Positive sentiment is more pronounced in popular movie taglines, emphasizing aspirational, uplifting, or exciting elements. This aligns with the promotional goal of creating an appealing first impression.

Less Popular Movies:

- **Tagline Sentiment Score:** 0.0264 (lowest across categories)
- Taglines for less popular movies display slightly lower positivity, often reflecting niche, subdued, or darker themes. This may limit their ability to captivate a broad audience.

Key Contrast: Popular movies strategically use positive language in taglines to offset the more negative tone of their overviews, creating a balanced emotional appeal. Conversely, less popular movies exhibit less pronounced positivity in taglines, which may contribute to weaker audience engagement.

4.3 Comparative Analysis of Overviews and Taglines

An interesting pattern emerges when comparing sentiments across overviews and taglines: they exhibit opposite trends. Overviews for popular movies tend to be more negative on average, using dramatic and emotionally intense language to engage audiences. In contrast, their taglines lean positive, often highlighting uplifting or aspirational elements. This complementary dynamic creates a narrative journey, starting with tension or conflict and resolving with a promise of hope or excitement. On the other hand, less popular movies show weaker contrasts between their overview and tagline sentiments. The less pronounced positivity in their taglines fails to offset the milder emotional tones in their overviews, leading to a less impactful overall emotional profile. This imbalance might explain the reduced engagement these movies experience.

4.4 Insights

1. **Emotional Intensity in Overviews:**
    a. Popular movies effectively use intense, dramatic language in overviews to convey high emotional stakes, capturing audience attention.
    b. Less popular movies tend toward more neutral tones, which may not evoke the same level of interest or anticipation.
2. **Strategic Positivity in Taglines:**
    a. Positive sentiment in popular movie taglines serves as a counterbalance to the negative tones in overviews, providing a sense of optimism or excitement that enhances audience appeal.
    b. Less popular movies fail to leverage taglines as a strategic tool, with lower positivity potentially underselling their stories.
3. **Balanced Emotional Appeal:**
    a. Popular movies demonstrate a purposeful interplay between negative overviews and positive taglines, creating a well-rounded emotional experience.
    b. Less popular movies miss the opportunity to use taglines to balance or elevate their narratives emotionally.

4.5 Conclusion

Sentiment analysis reveals a strategic emotional approach in popular movies, where the interplay between intense overviews and uplifting taglines fosters broad appeal. Less popular movies, with their weaker contrasts and milder emotional tones, lack this dynamic, potentially limiting their ability to captivate diverse audiences. These findings highlight the value of crafting complementary emotional tones in storytelling and marketing to maximize engagement.

5. Conclusion

This report provides a comprehensive analysis of the relationship between textual features, sentiment trends, and movie popularity, offering actionable insights for creators and marketers. Key findings reveal that popular movies effectively utilize dynamic, emotionally resonant language and universal themes in both storytelling and marketing materials, setting them apart from less popular counterparts. Through detailed word frequency, n-gram, and sentiment analyses, the study highlights the linguistic strategies that contribute to audience engagement.

Key Takeaways:

1. **Dynamic and Universal Themes:** Popular movies consistently emphasize dynamic narratives, emotional depth, and universally appealing themes such as love, resilience, and transformation. These elements are conveyed through action-oriented and emotionally charged language in overviews and taglines, enhancing their broad relatability.
2. **Strategic Marketing Language:** Taglines in popular movies strike a balance between optimism and excitement, complementing the intense, dramatic tone of their overviews. This interplay creates a well-rounded emotional experience that resonates with audiences.
3. **Contrast with Less Popular Movies:** While less popular movies often explore similar themes, their execution lacks the intensity, breadth, and strategic emotional balance that define popular films. Niche or darker content and weaker engagement in taglines may limit their broader appeal.
4. **Sentiment as a Driver of Appeal:** Sentiment analysis reveals that popular movies leverage negative tones in overviews to convey high stakes and dramatic tension, paired with positive tones in taglines to provide resolution and excitement. This emotional journey is less evident in less popular movies, which may account for their reduced engagement.

Implications for Stakeholders:

- **Content Creators:** Incorporating universal themes, diverse representation, and dynamic narratives can enhance a movie’s audience appeal.
- **Marketers:** Optimizing taglines with emotionally engaging and aspirational language can maximize promotional impact and broaden reach.
- **Industry Strategists:** The strategic use of language in both storytelling and marketing is crucial for capturing and sustaining audience interest.

In conclusion, this report underscores the importance of thoughtful linguistic and emotional strategies in defining a movie’s success. By aligning storytelling and promotional efforts with audience preferences, creators and marketers can craft compelling narratives and campaigns that resonate across diverse demographics, ensuring stronger engagement and popularity.

6. Case Study: Sentiment of Top 20 Most Popular Movies

This sentiment analysis explores the emotional tone of overviews and taglines for the top 20 most popular movies. The findings reveal distinct patterns in how these elements are crafted to resonate with audiences.

1. Overview Sentiments

The overview sentiments exhibit significant variation, reflecting the diverse emotional tones of these films:

- **Examples of Negative Sentiments:**
    - *Se7en*: -0.9825
    - *The Dark Knight*: -0.9485
    - *Joker*: -0.9337
    - *The Godfather*: -0.9371 These movies often explore intense, morally complex, or dark themes, aligning with their negative sentiment scores.
- **Examples of Positive Sentiments:**
    - *Parasite*: 0.8807
    - *Django Unchained*: 0.5574
    - *The Matrix*: 0.3612 These films often present narratives with uplifting, transformative, or aspirational elements, contributing to their positive sentiment.

The prevalence of negative sentiments in overviews highlights the popularity of movies with emotionally intense and dramatic storytelling. These narratives captivate audiences by delving into themes of struggle, conflict, and transformation.

2. Tagline Sentiments

Taglines tend to skew more neutral or slightly positive, with notable examples of extremes:

- **Examples of Positive Sentiments:**
    - *Schindler's List*: 0.7506

    - *The Green Mile*: 0.5859
    - *Joker*: 0.5719
    - *Django Unchained*: 0.5267 These taglines evoke hope, curiosity, or inspiration, engaging audiences through emotional or aspirational hooks.
- **Examples of Negative Sentiments:**
    - *Se7en*: -0.8176
    - *Inception*: -0.5423
    - *The Lord of the Rings: The Return of the King*: -0.5423 These taglines mirror the darker tones of their narratives, reinforcing their thematic depth.

Popular movies often balance dark overviews with taglines that are either positive or neutral, creating a contrast that captures audience attention. When both sentiments are negative (e.g., *Se7en*), the films fully embrace their darker themes.

3. Contrasting Sentiments

Several movies demonstrate a strategic contrast between overview and tagline sentiment:

- *Joker*: Overview (-0.9337), Tagline (0.5719)

The tagline provides a hopeful and intriguing counterbalance to the film's dark narrative.

- *The Dark Knight*: Overview (-0.9485), Tagline (0.4588)

The tagline's positivity contrasts with the intense themes of the overview.

- *Schindler's List*: Overview (0.1779), Tagline (0.7506)

Both sentiments lean positive, emphasizing the movie's uplifting message despite its heavy subject matter.

This contrast between overview and tagline sentiment often reflects a deliberate strategy to balance emotional intensity with intrigue or hope. It can enhance audience appeal by presenting a multifaceted emotional experience.

Conclusion

1. Negative Sentiments Dominate Overviews:

Dark, intense themes are a hallmark of popular movies, as reflected in their negative overview sentiments. These themes resonate strongly with audiences seeking compelling and emotionally charged narratives.

2. Positive Sentiments Dominate Taglines:

Taglines often offset dark overviews with positive or neutral sentiments, creating intrigue or a sense of hope that draws viewers in.

3. Contrast as a Strategy:

Strategically contrasting sentiments between overviews and taglines enhances emotional complexity and marketing appeal. Films like *Joker* and *The Dark Knight* exemplify this approach, capturing audience attention with their layered emotional tones.

This analysis underscores the importance of sentiment as a tool for storytelling and marketing, revealing how emotional tone can shape audience engagement and movie popularity.