College of Engineering and Information Technology

Department of Information Technology

Major: Data Analytics

Data Visualization course Project

Comprehensive Analysis and Visualization of Movie Datasets

Prepared by:

Mustafa T M FarajAllah – 202310819

Supervised by:

Dr. Salam Fraihat

Academic Year 2024- 2025 – Spring

Abstract—This project focuses on a comprehensive analysis and visualization of 6 movie datasets, performing data preparation, and exploratory data analysis (EDA). The workflow begins with data preprocessing to ensure data quality and consistency. EDA techniques are then employed to uncover key trends, patterns, and insights within the datasets, including relationships between revenues, budgets, genres, and other key variables. The objective of this project is to make data-driven storytelling that can help with decision-making and provide insights/tips for movie production companies. Any entertainment institution that decides to follow the hidden trends discovered can expect an increase in its overall revenue by 25%.

Introduction:

Movie production is one of the fastest growing industries that combines art, business, and culture to entertain, educate, and create trends. They are typically shown in theaters, on television, or through streaming platforms, and have become a global language that connect people across borders, affecting their perspective on the world as movies raise awareness. What makes movies special is that it can gain revenue from multiple sources both during and after the movie was released.

In this project, we aim to analyze the characteristics of movies to extract insights and trends from the data in the form of visualizations that can be presented to stakeholders, including businesses and investors. Those graphs would help in understanding the market and take big decisions safely with minimal risk and achieve the best potential revenue.

Finally, we will construct a machine learning model trained on the data using different algorithms, then choose the best one in terms of predicting the movie revenue post-release.

Objectives:

- Creating consistent visualizations that tell a story.
- Help in decision-making for the top movie production companies.
- Provide insights/tips for relatively small movie production companies to increase their revenue.
- Build a machine learning model that can predict the movie overall revenue after release.

Our criteria:

- **Know Your Data (KYD):** Have deep understanding of our datasets by being aware of key features.
- **Consistency:** Apply the necessary cleansing and preprocessing on the data to ensure no bias is introduced into any graph.
- **Transparency:** Using the original data with no manipulations to change an existing trend.
- **Simplicity:** Visualize the data using the appropriate graphs, making sure that the graph reaches the stakeholders clearly with no complexity.

- **Generalization:** Ensure that the machine learning model can perform well on unseen data after learning from the training set, making sure that the model is not overfitting nor memorizing the data trained on.

The datasets used:

To achieve our objectives, we searched the internet for valid, real-world datasets. We stumbled abone the following 6 datasets:

1. The Numbers dataset (not clean from the author):
   The Numbers is the most reliable website that calculates the movie revenue in real time. The website is constantly being updated, proving its reliability. Additionally, there are many tabs in the website to gain important data from. A Kaggler scrapped a tab that contain financial information regarding some movies, around 6 thousands records and downloaded the data on Kaggle.
   The issue with this dataset is that it contains a lot of empty cells (NaNs).

2. TMDB dataset (not cleansed from the author):
   TMDB is the biggest movie dataset in the world, containing hundreds of thousands of movies. A Kaggler scrapped their website and downloaded the data on Kaggle.
   However, this dataset is unreliable, as existing movies in the world did not exceed 500 thousands movies, while the data has over 1.2 million movies. A question raised, why use it if it is not reliable? Well, as we discussed in dataset number 1, that dataset contains a lot of empty cells (NaNs), we wanted to fill those missing cells using real-world data, which TMDB can do. But we just mentioned that the data is not reliable? In this case, we will only take the records in TMDB that correspond to the The Numbers dataset, as we know for sure that The Numbers dataset is reliable.

3. Box Office MOJO dataset (not cleansed from the author):
   Box Office MOJO is a reliable website that stores Box Office revenue for movies. They have a tab that sums all Box Office revenue in a certain year. A Kaggler scrapped their website and downloaded the data on Kaggle.
   However, this dataset only calculates the Box Office revenue of the top 200 movies, this dataset alone is insufficient to capture global trends, so we had to merge it with more datasets to be able to capture real-world trends.

4. Box Office August 2019 (cleansed from the author):

   After searching for another Box Office dataset for long time, we stumbled abone a GitHub repository that had a diverse number of datasets, among them was Box Office August 2019. We did not find a source for that dataset however, the records in it actually match the real-world output, not to mention that no reports were issued for that repository proving its reliability. The only issue with this data is that it is not updated, as it contain records up to August of 2019.

5. Netflix Stock Market dataset (cleansed from the author):

   We kept searching for Netflix Revenue dataset but did not find one. So, we thought of employing Netflix Stock Market data to mimic the revenue of the organization. The data was found on Kaggle.

6. IMDB (manual entries):

   We searched the web and spotted that IMDB has a record in their website for every movie.
   The website counts the number of movies produced in each decade, we created a 4 record dataframe manually using the website data.

Data Preparation:

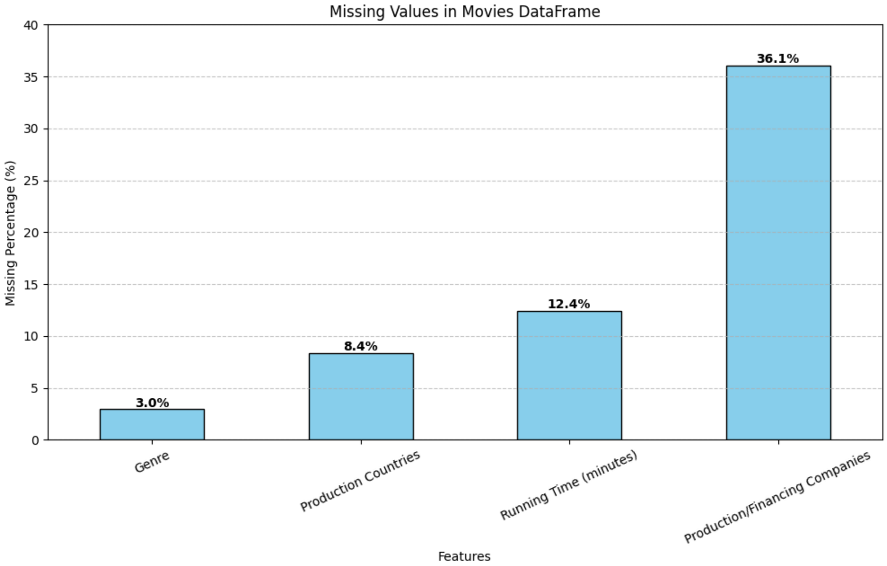
In this phase, we mainly performed the following steps for only the uncleansed datasets:

- Data understanding and cleansing
- Data Merging (if needed)
- Feature engineering

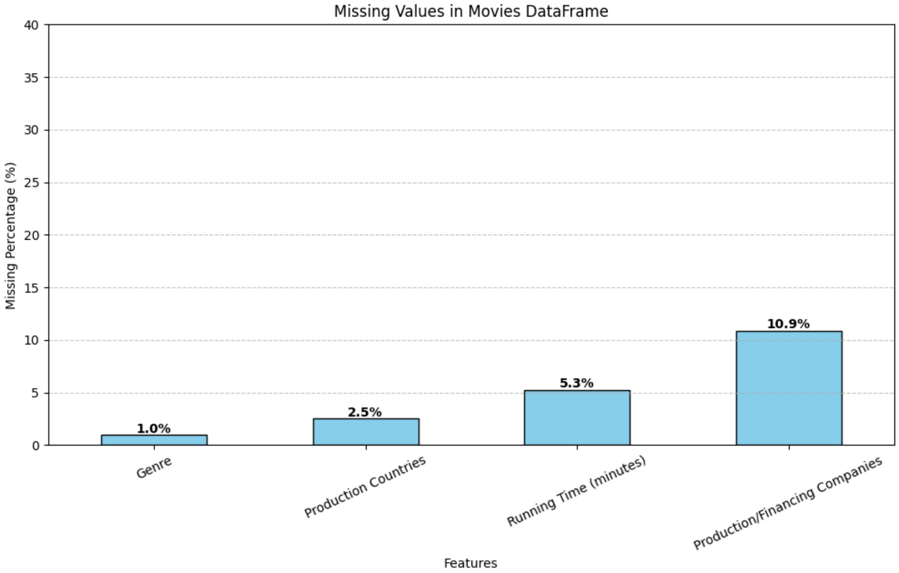
In Data understanding and cleansing step, we performed Exploratory Data Analysis (EDA) to under the data and its issues, such as duplicated records, percentage of null values, and cleaning the features that were inconsistent or misleading.

In Data Merging step, we imputed the missing values in “The Numbers” dataset with the corresponding records in the “TMDB” dataset. Since we did not have a primary feature to merge on, this step was challenging. At first, we thought that the movie title was unique however, turns out this feature contained duplicated entries, and it was because of “movie remake” purposes. So, we decided to merge with 2 features, movie title and release date, but turns out that the release date was inconsistent in both datasets (e.g. if we tried to merge the same movie “Robin Hood”, the date in The Numbers dataset does not match the date in the TMDB dataset). To fix this issue, we used the .merge_asof() function offered from pandas, which had the ability to rewind back in time to find the closest match

Sample of 4 features before merging “The Numbers” with the “TMDB” datasets:



The same sample of features after merging “The Numbers” with the “TMDB” datasets:



The “Box Office MOJO” dataset was easily merged with “Box Office August 2019” and “The Numbers” datasets (we merged the Box Office MOJO with “The Numbers” because “The Numbers” had Box Office features).

In Feature engineering step, we created some features out of our data, those features were for visualization purposes.

After completing those steps, we saved the cleaned and enhanced version of the datasets to save computation resources and time.

Data preprocessing:

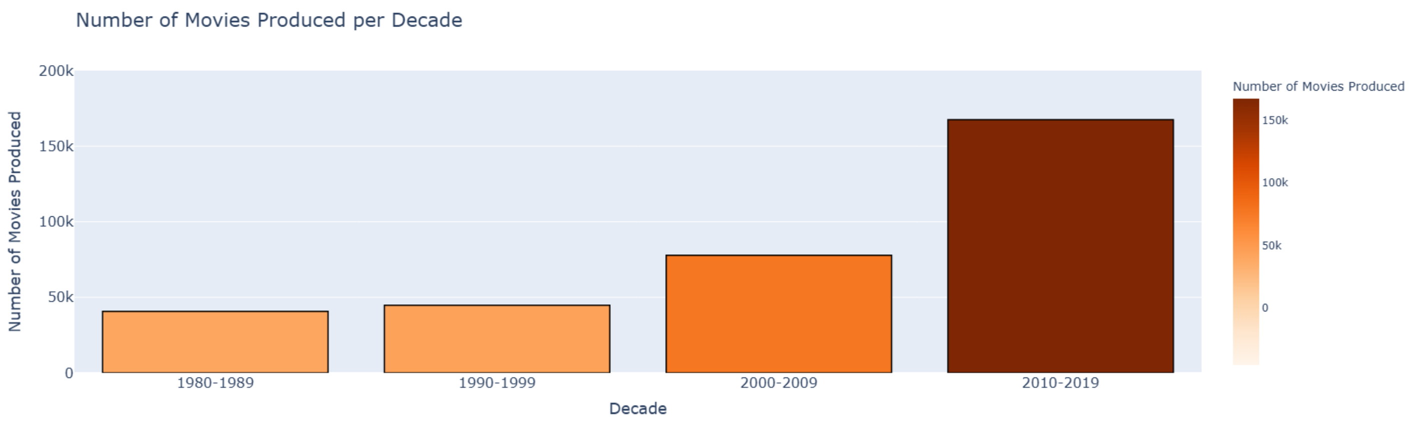
In this phase, we imputed the missing values in "The Numbers" dataset with the mean and median depending on the distribution of the specific feature and dropped 4 features "Est. Domestic Blu-ray Sales (USD)", "Est. Domestic DVD Sales (USD)", "Total Est. Domestic Video Sales (USD)" and "theater_ratio" as they were filled with empty cells (NaNs). This phase was only applied to "The Numbers" dataset as it was our primary visualization resource and because we wanted to create consistent graphs.

Data Visualization:

In this phase, we will analyze the cleaned movie datasets and extract insights, patterns, and meaningful relationships. Additionally, we used the following applications to produce proper visualizations that matches our criteria:

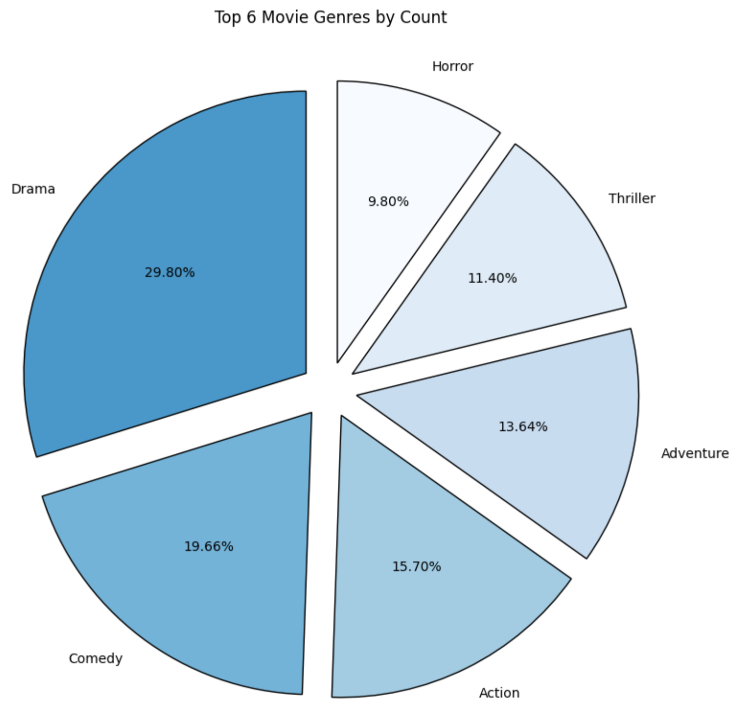
- Python: Using code, we filter and re-group the data to produce the imagined graph.
- Microsoft Power PI: Using its interactive interface and data visualization capabilities, we construct the imagined graph by connecting to the processed data and configuring visual elements to represent filtered and re-grouped information.
- Microsoft Excel: Using this applications, we were able to introduce some amazing graphs that were hard to produce with the previous applications.

Question one) Are movies a growing industry?

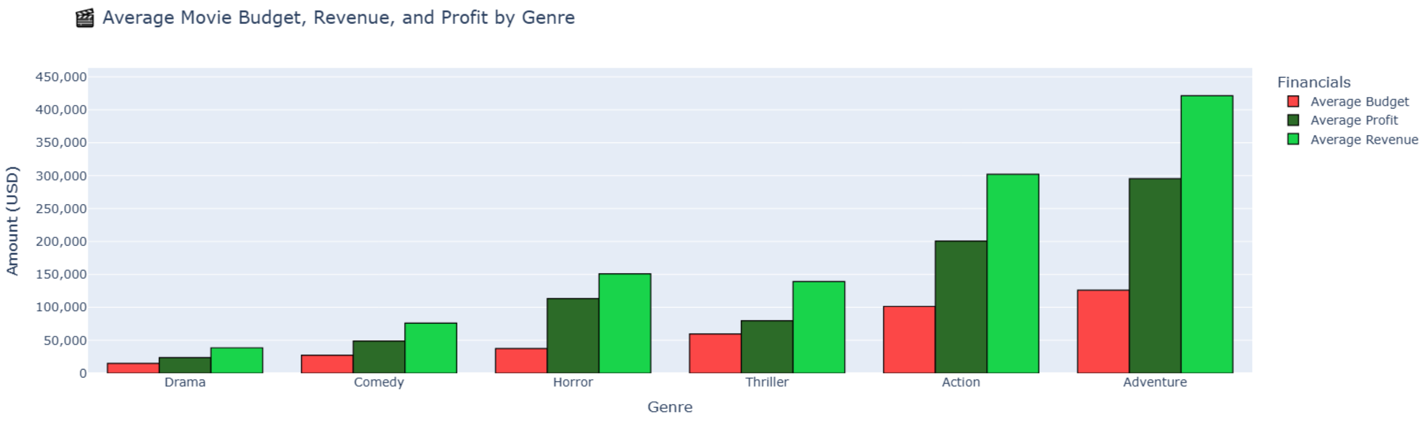


We can see clearly that the movie industry is a growing industry, with thousands of movies produced every year. This trend also indicates that the movie industry is getting more competitive, meaning that as a production company, you must understand the audience preferences to ensure movie success.

Question two) What are the Most Popular Movie Genres, and why is it popular?

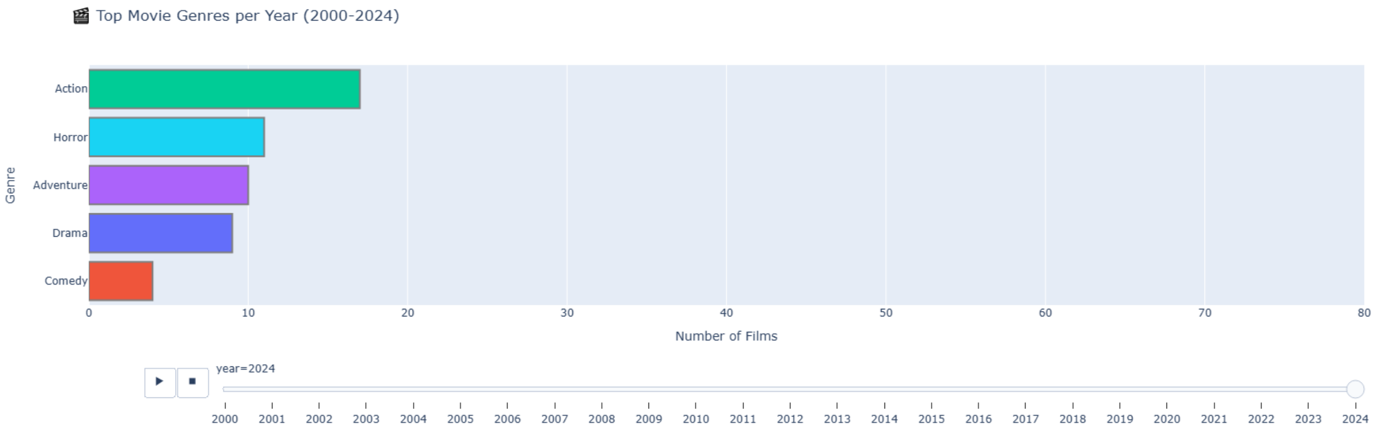


Drama genre is the most common, making up nearly 30% of the dataset. Why might that be the case?



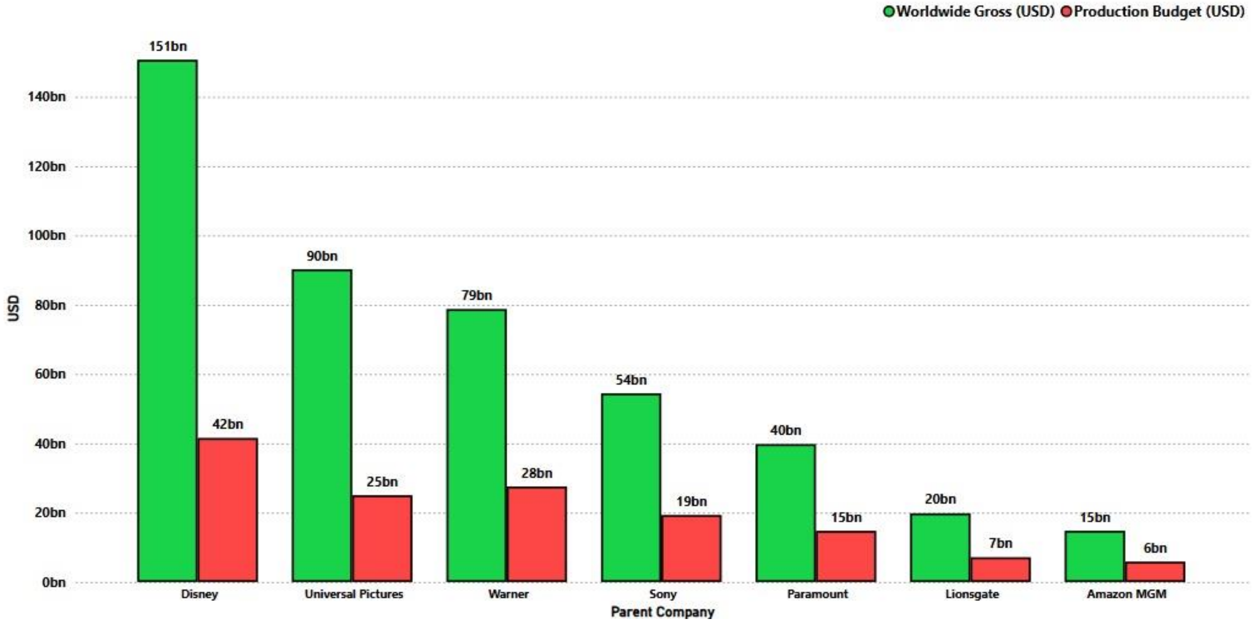
A possible reason why Drama genre is the most common, is that it requires the smallest movie production budget out of all of them, making it an excellent choice for startup movie production companies as it is cheap to create and does not require a lot of money commitment. Thriller genre might not be the right choice as it does not make enough Return On Investment (ROI) compared with the other genres. Action and Adventure genres are the most expensive to produce, moreover, they are the most profitable choices.

Question three) What is the current trending genre?


🎬 Top Movie Genres per Year (2000-2024)

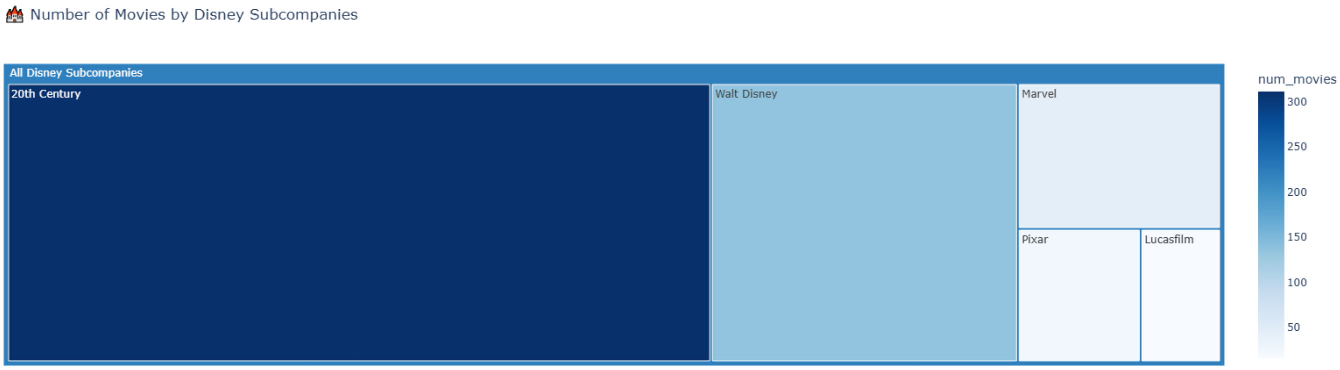
It is important to mention that movies usually take minimum of 2 years in production, meaning that we must consider multiple prior years before committing in creating a movie. Action and Adventure genres seems to be the trending genres for the years 2022, 2023, and 2024.

Question four) What are the top movie production companies?

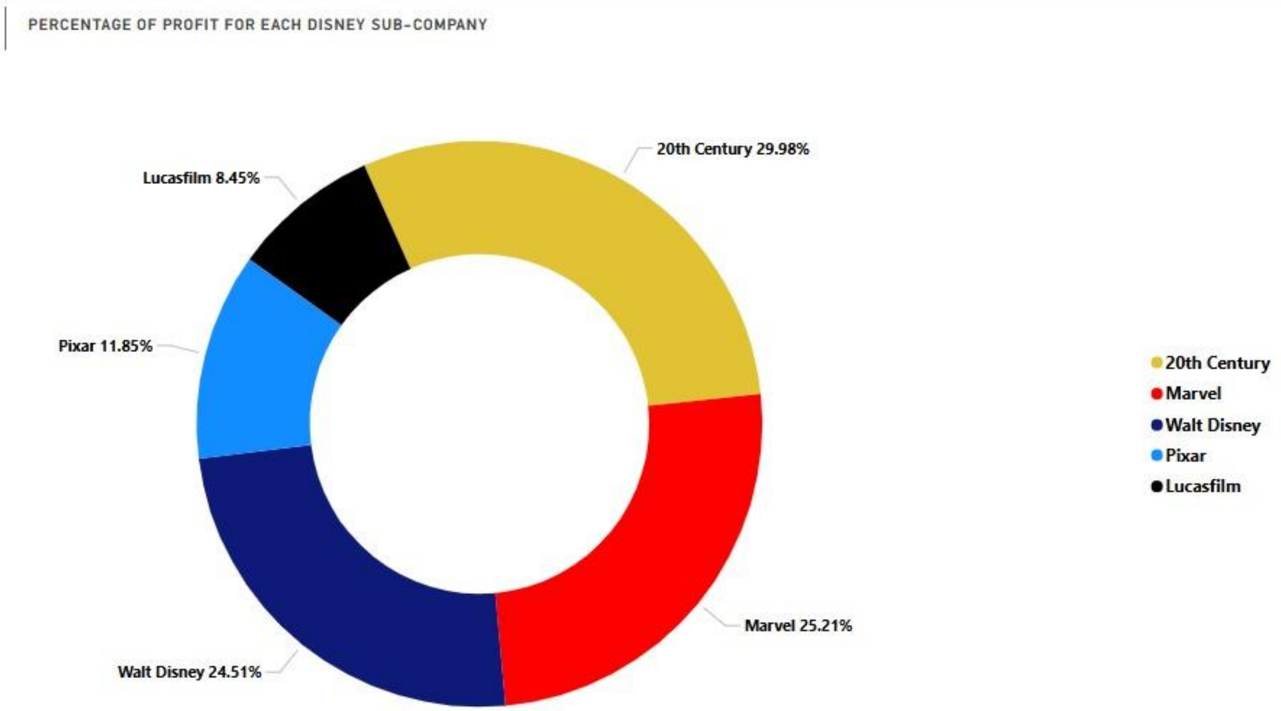


This chart highlights the financial performance of major film studios over several decades, comparing their total production budgets with total revenues. Disney clearly leads the industry, with the highest revenue and budget, reflecting its dominance. Universal Pictures and Warner. follow closely, showing strong returns. Sony, Paramount, and Lionsgate operate on smaller budgets but still generate significant revenue, with Lionsgate showing efficient returns despite its small scale.

Since Disney is the biggest movie production company and has a lot of subcompanies under it, what is the number of movies generated from each subcompany, does those subcompanies return similar profit rates?



We can observe that the leading subcompany in terms of movie releases is “20th Century”, producing over 300 movies. Coming up are “Walt Disney” with 136 movies, “Marvel” with 47 movies, “Pixar” with 26 movies, and “Lucasfilms” with 17 movies.



Although “Marvel” has produced fewer movies (47 movies) compared to “20th Century” and “Walt Disney”, it stands out as one of the efficient studios of Disney, delivering the second highest profit. “Pixar” and “Lucasfilms” also perform exceptionally well. Their profit is less compared to other subcompanies however, they have the smallest number of movies, with “Pixar” produced 26 movies and “Lucasfilms” 17 movies. Walt Disney (main studio) contributes a large chunk, but its efficiency is less, likely because of the mix of film types. “20th Century” has biggest profit returns but also has the largest number of releases suggesting quantity over quality issue.

Question five) Do top movie studios tend to produce more adult oriented or child friendly content?

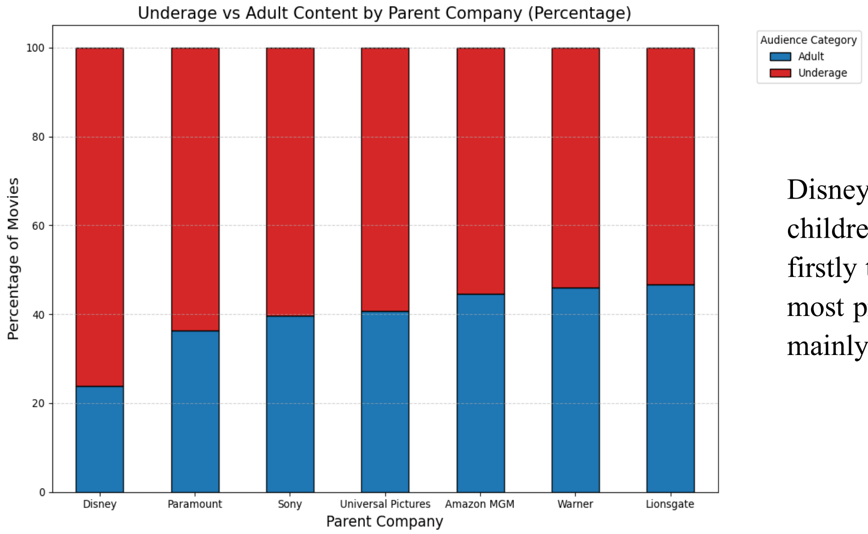
In this question, we used the feature “MPAA Rating” which in simple words, a film rating system. This feature had a lot of values under it including ['G' - 'R' - 'PG-13' - 'PG' - 'M/PG' - 'NC-17' - 'Unknown' - 'Not Rated' - 'Open']. To answer this question, we had to aggregate those features into two groups, Adult, and Underage:

- For Underage (under 17): 'G', 'PG', 'PG-13'
- For Adults (18 and above): 'R', 'NC-17', 'M/PG', 'Not Rated'

A question might rise, why will ‘M/PG’ and ‘Not Rated’ be classified for adults?

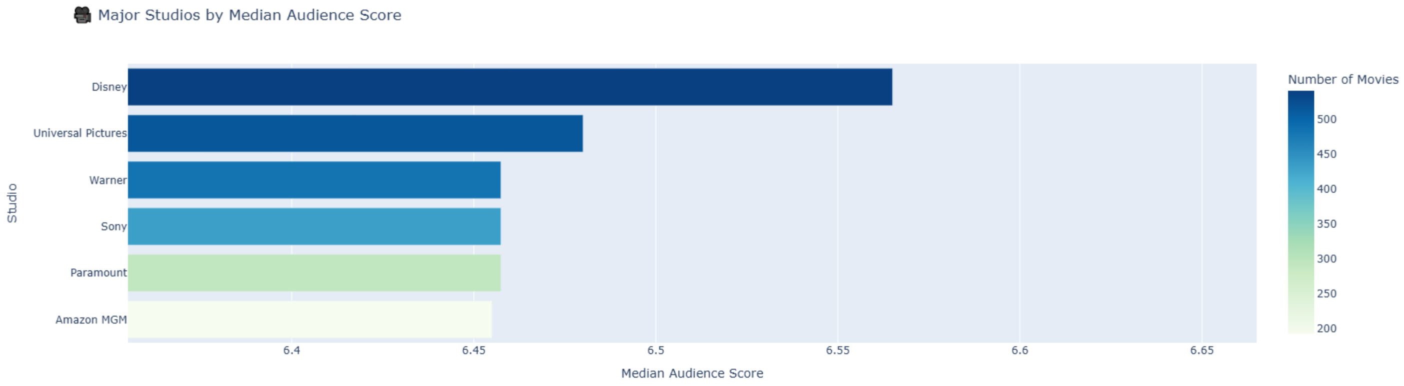
- 'M/PG' means Mature, and most countries set Mature age to 18, meaning those movies are for adults.
- 'Not Rated' is a bit controversial, as in one hand, some production companies cannot afford to go through the MPAA test, while other companies intentionally do not rate their movies to make the audience hooked.

The 'Unknown' was discarded from this analysis.



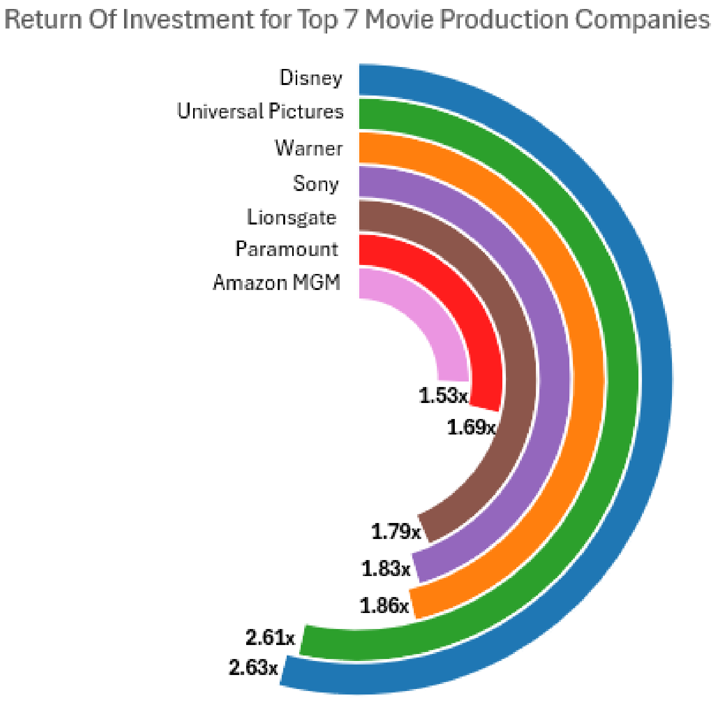
Disney seems to produce content oriented for children more than to adults. Which tells us firstly that movies created for children are the most profitable of all and secondly, Disney is mainly kids movie production company.

Question six) What are the median vote scores for the top movie production studios?

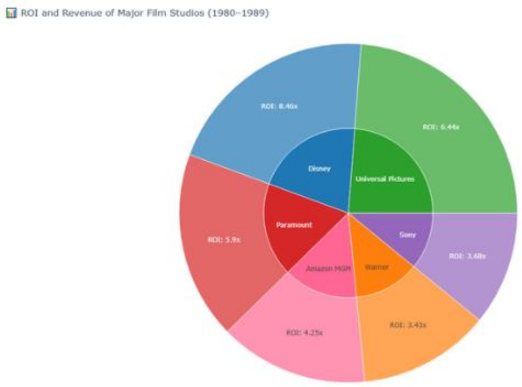


Does this chart contribute with the companies revenue? Yes it does, remember the chart in "Question four", We can see that the order of bars is like the order in "Question four", showing that indeed, revenue corresponds to the audience rating score. Additionally, we can observe that since Disney target audience is children, parents must like Disney content, giving very high median rating.

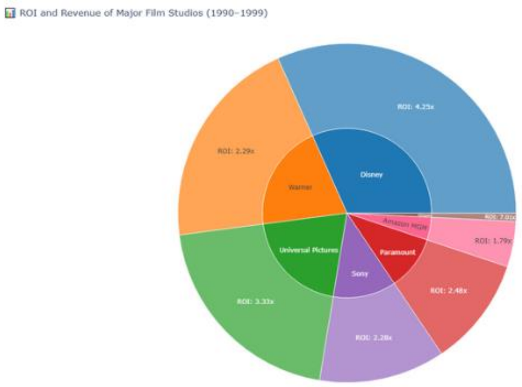
Question seven) What is the Return Of Investment (ROI) for the biggest movie production companies?



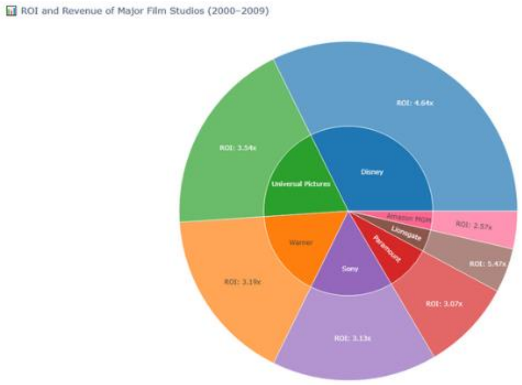
We can see the ROI for all top production companies, with the top companies being "Disney" and "Universal Pictures". But across different decades, what was the highest ROI company?



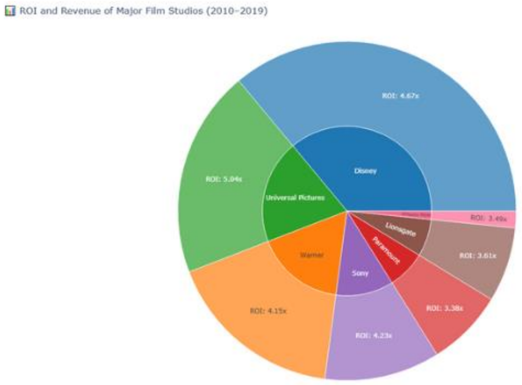
📊 ROI and Revenue of Major Film Studios (1980–1989)



📊 ROI and Revenue of Major Film Studios (1990–1999)



📊 ROI and Revenue of Major Film Studios (2000–2009)

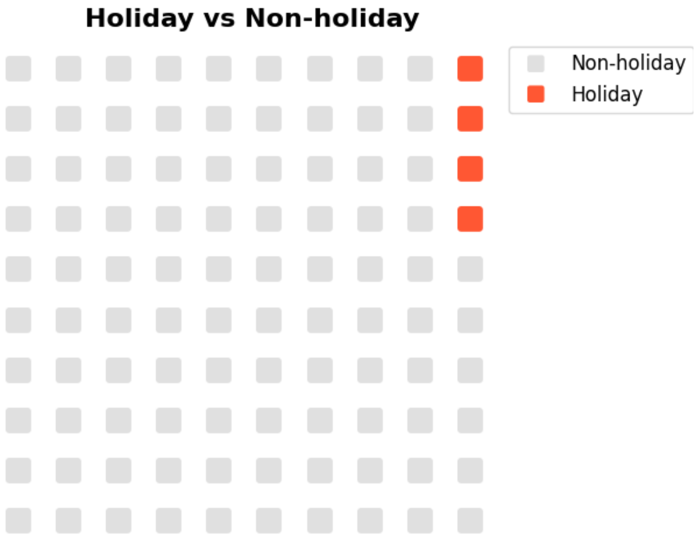


📊 ROI and Revenue of Major Film Studios (2010–2019)

"Disney", "Universal Pictures", and "Warner" are the biggest competitors across different decades.

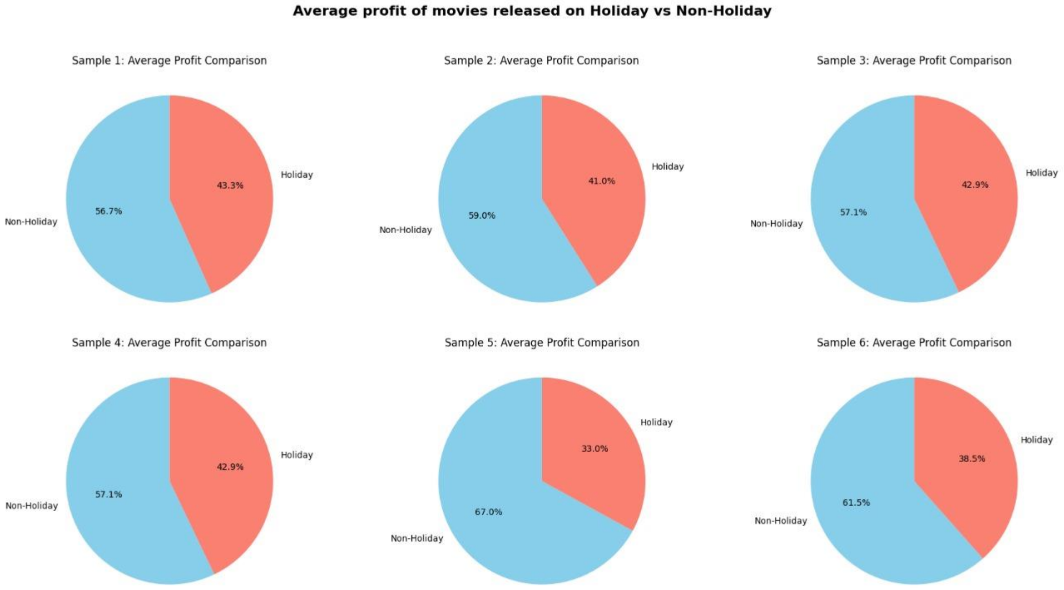
Question eight) Should companies release movies during official holiday days or not?

To answer this question, we employed a library called "Holidays". This library generates holidays based on the country and date. A problem we faced is imbalance between non-holiday and holiday days, under the same feature "is_holiday", as the number of holiday movies was just over 250 movies, and the non-holiday was over 5000 movies.



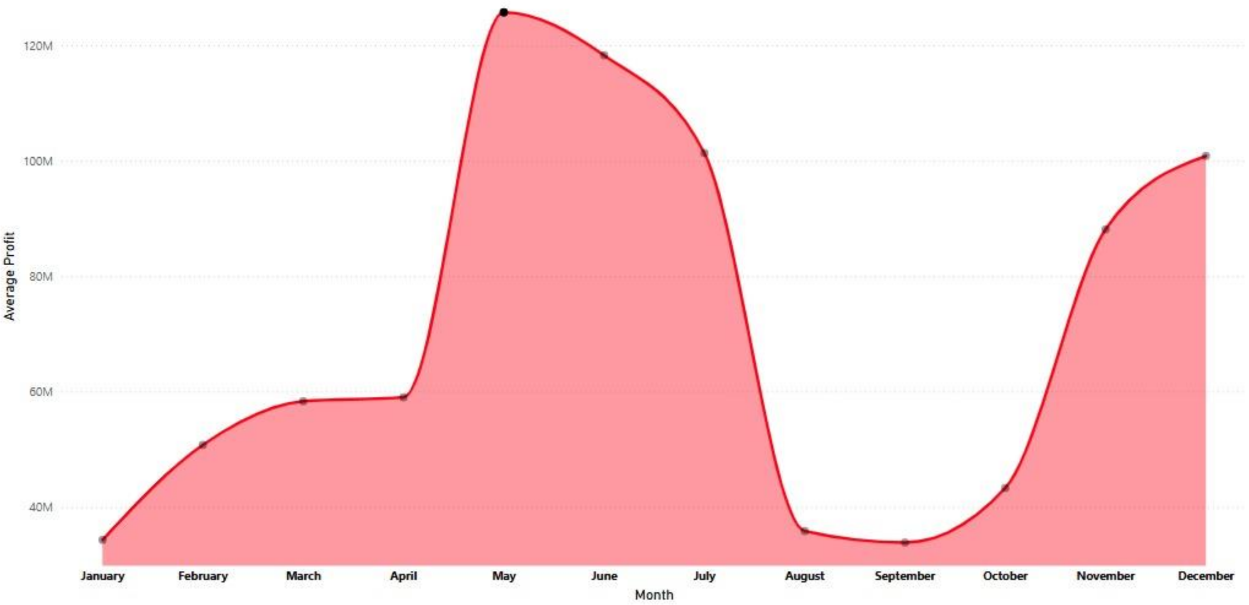
Looking at the Waffle chart, we can observe the issue we discussed above under the question. The feature distribution is imbalanced and comparing them with no extra procuration is not fair.

To solve this issue, we will be taking 6 random record samples that contain the class Holiday, and compare each subset with the Non-holiday profits.

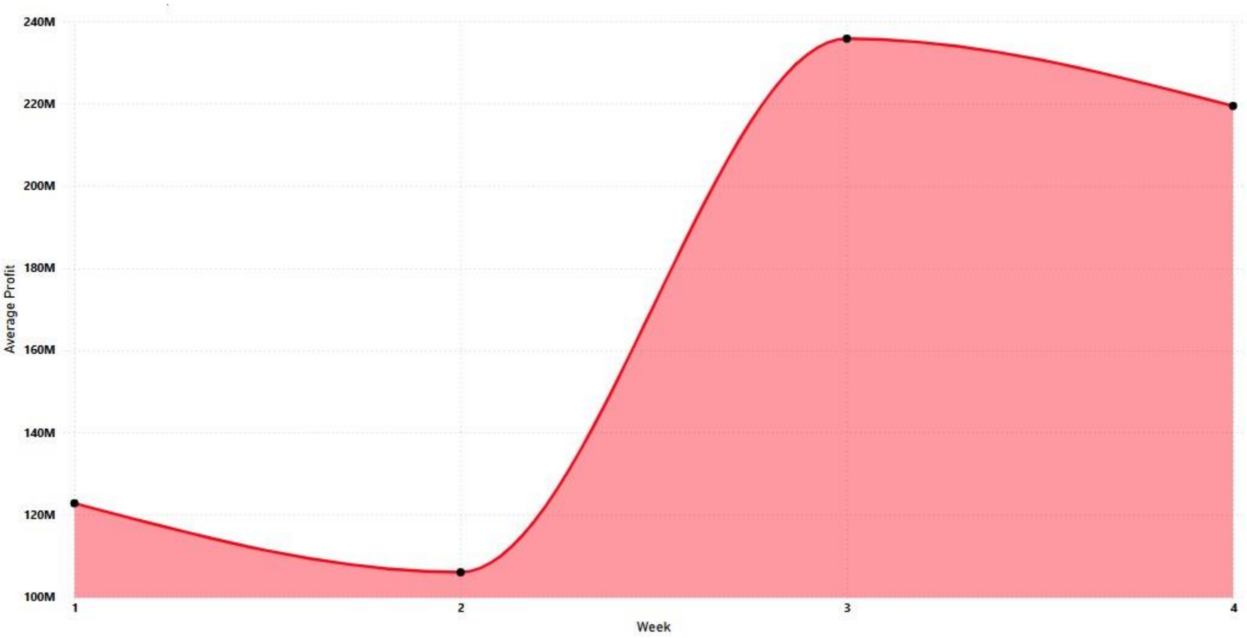


Notice how releasing a movie during the holiday will significantly reduce the movie profits, which is logical. During **News year,** people usually tend to go outside for public and famous places, but not for movie theaters.

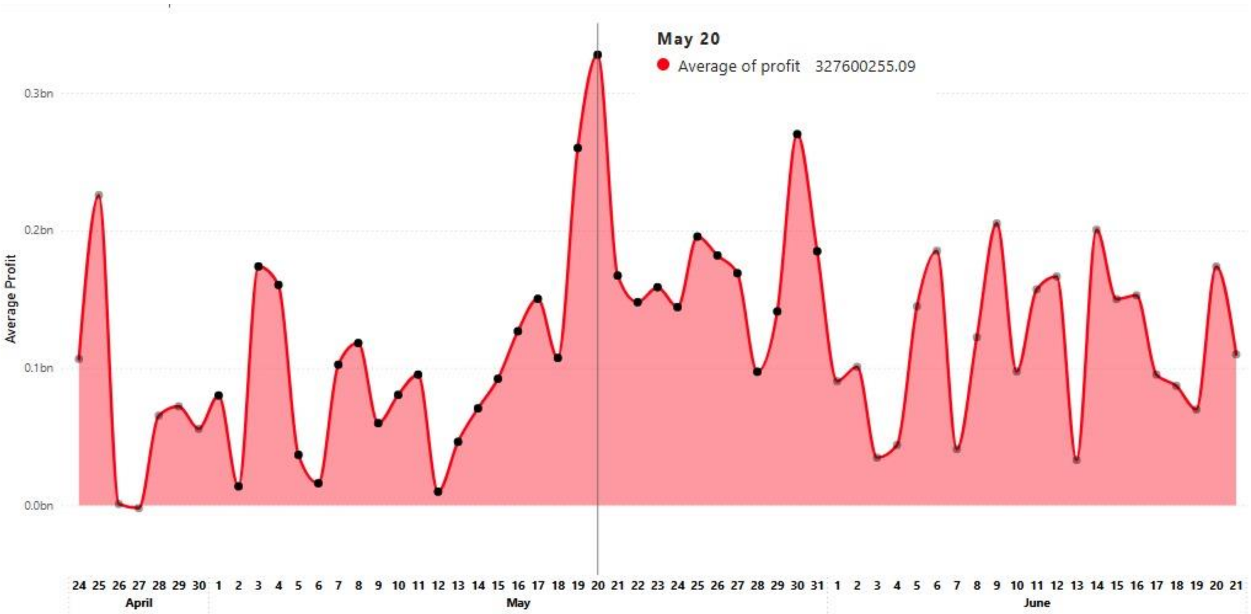
Question nine) What is the most profitable month and day to release a movie on?



Notice May, June, and July are the most profitable months to release a movie on. However, May is the most profitable one of them.



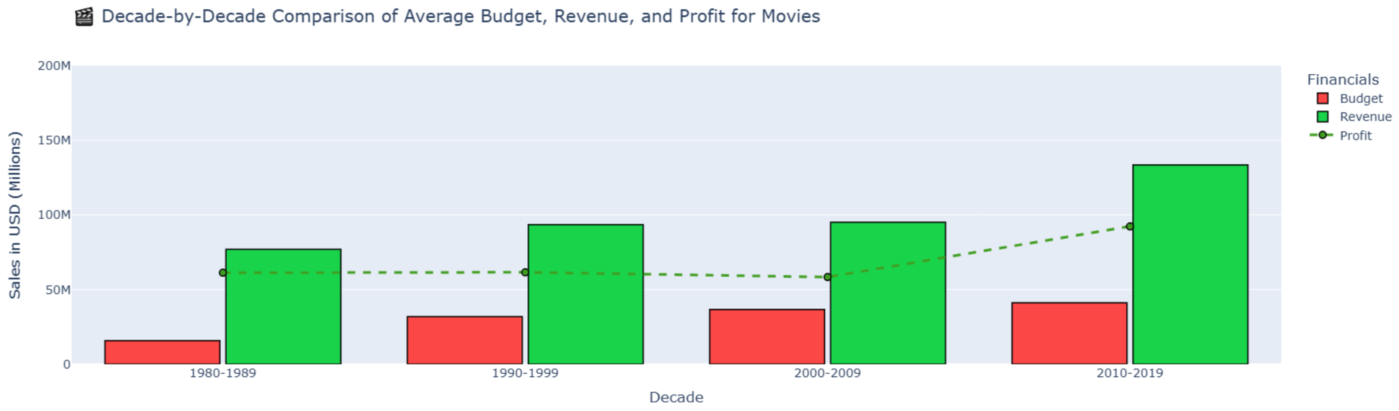
We drilled down in May, and noticed that releasing on the third or fourth week will result in highest returns.



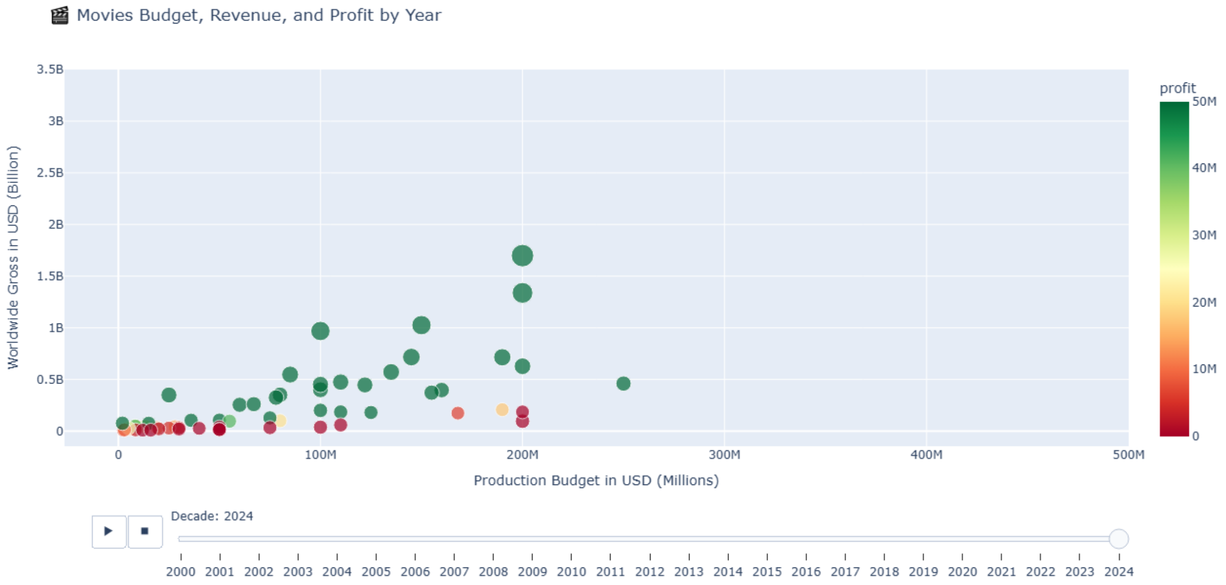
To be exact, the 20th is the most profitable day to release a movie on.

Question ten) In general, does the movies budget and revenue increase with time? Is it a profitable relationship?

Answering this question is important, as we want to know whether increasing the budget would increase the revenue resulting in a higher profit margin.



While the revenue was increasing, the budget does not seem to grow as much, suggesting a low positive correlation between the budget and revenue.

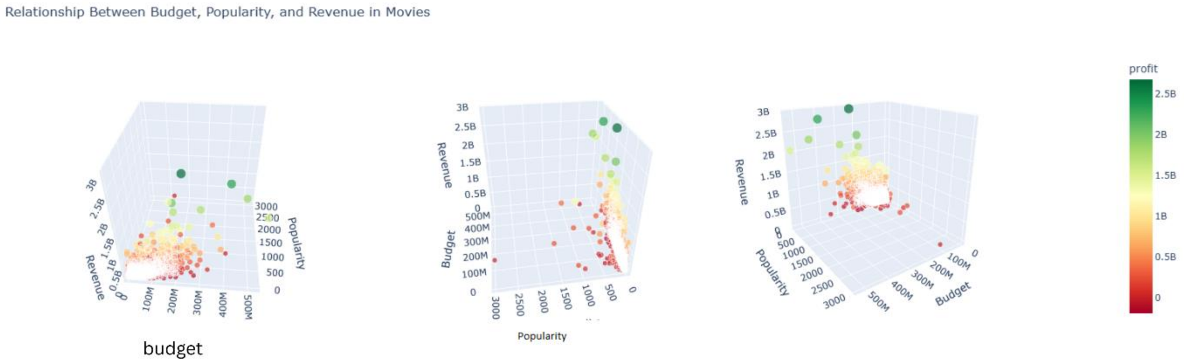


We read a study that found a relationship between a movie's budget (independent variable) and its revenue (dependent variable). While budget plays a role in predicting revenue, the study suggests that other factors, such as number of actors also contribute to a movie's financial success. This study reinforces our observation that higher budgets tend to correlate with higher revenue, though the relationship is influenced by additional factors. Notably, in the lower right section of the visualization (marked by red and yellow bubbles), we can see outliers, movies with exceptionally high production budgets but low revenue. These cases suggest that while a large budget can contribute to success, it does not guarantee profitability, further emphasizing the complexity of revenue prediction in the film industry.

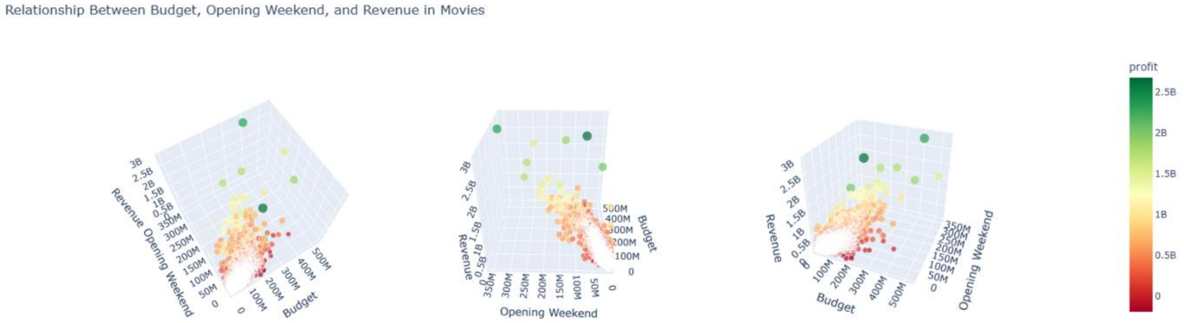
Question eleven) What other features can we add to increase the correlation with budget and revenue?

Firstly, we want to test with the feature “popularity”. This feature indicates how people were motivated to watch the movie pre-release. Will this feature create a strong connection with the budget and revenue?

Looking at the chart below, when popularity increases, it does not really mean that the revenue would increase nor budget, which suggest a very weak correlation, if it even exit.


Relationship Between Budget, Popularity, and Revenue in Movies

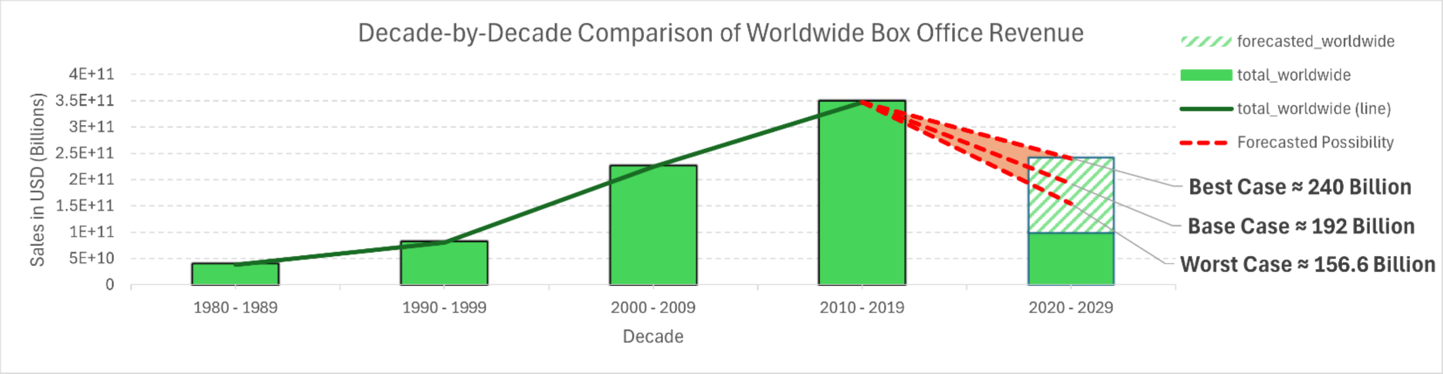
How about the “Opening Weekend (USD)”, does it affect the connection with budget and revenue? is it a positive relationship?


Relationship Between Budget, Opening Weekend, and Revenue in Movies

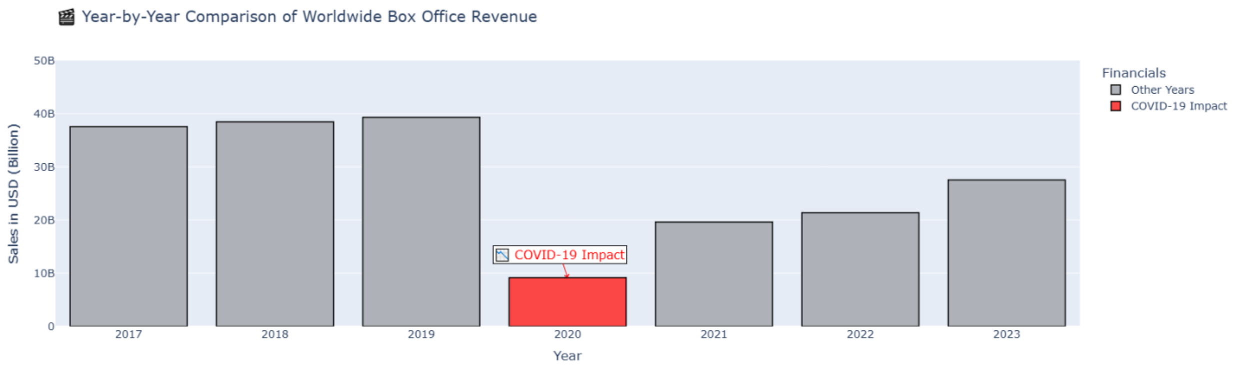
Looking at the above chart, we can observe that “Opening Weekend (USD)” has a positive correlation with budget and revenue. Which indicates that if the movie was released in theaters, and got in the first week good revenue, then the overall revenue of movie will be positive. Additionally, this feature will be valuable for us in the machine learning phase.

Question twelve) After learning that movies on average are profitable, what about Box Offices?

Box offices are a main source of income for movies, and since movies generate significant revenue through box office earnings, one might initially assume that box offices themselves are inherently profitable, however, the profitability of a movie and the profitability of the box office are distinct concepts influenced by many factors.



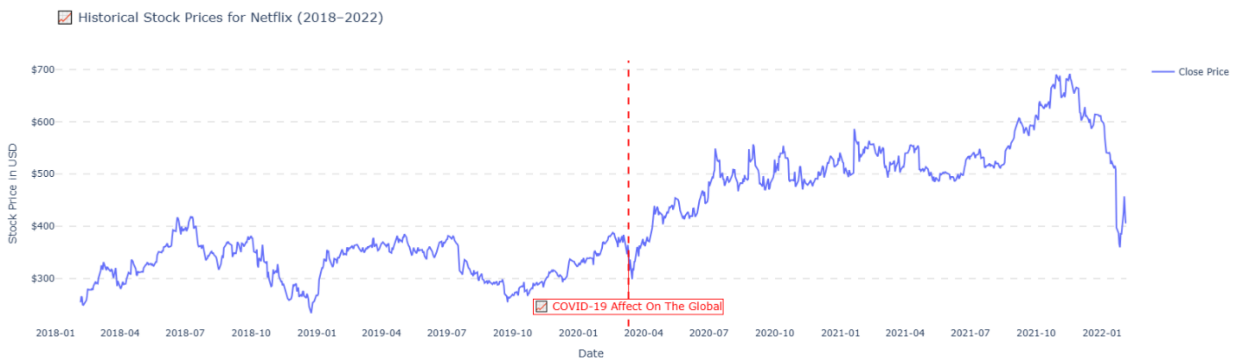
Analyzing the chart, we can see that box offices were gaining billions of dollars decade after decade, however, for the current decade box offices are forecasted to gain lower revenue than the previous decade. Such a dramatic loss indicates a global event that occurred.



After drilling down, we can see that 2020 was the worst performing year, because COVID-19 hit. This caused theaters to shut and delay major movie releases. Box office revenue dropped to historic lows as people stayed home. When the pandemic became controlled, lock-down was lifted, theaters reopened, and big movies returned to the big screen. Audiences slowly came back, and by the mid-2020s, the box office was recovering. However, why did box offices still not gain their previous world record revenues?

Question thirteen) How streaming websites affected box office revenue?

When COVID-19 hit, most businesses shut down or lost a large percentage of their revenue. Surprisingly, this crisis somehow created competitors for box offices, such as Netflix and other movie streaming websites. When COVID-19 hit, most businesses shut down or lost a large percentage of its revenue. Surprisingly, this crisis somehow created competitors for box offices, such as Netflix and other movie streaming websites.



When COVID-19 was announced as a global pandemic, all companies around the world drop in stock market, Netflix included. However, Netflix quickly bounced back as more people turned to streaming while staying at home during lockdowns. We can clearly see the huge increase in Netflix stocks just short period after the pandemic was announced. Notably, in 2022, Netflix lost a large chunk of its stock, and that’s because they raised their prices losing over 200,000 subscribers.