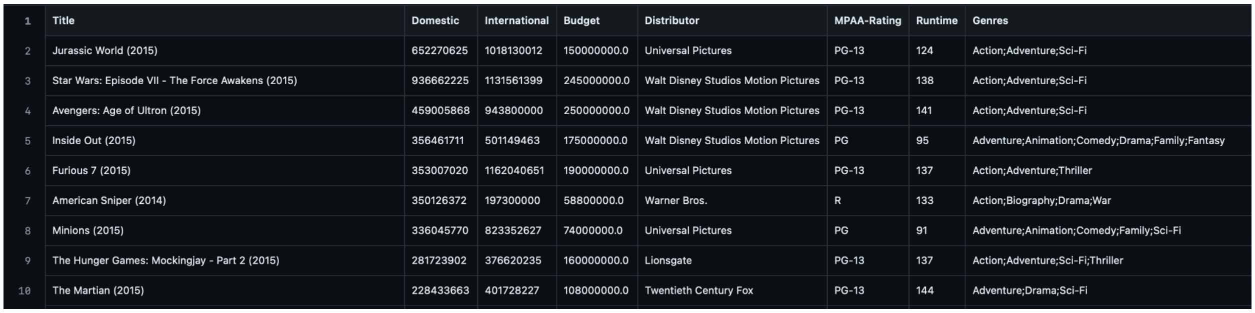
Samrawit Basazinew
James Lee
Jessica Price
Andrew Vick

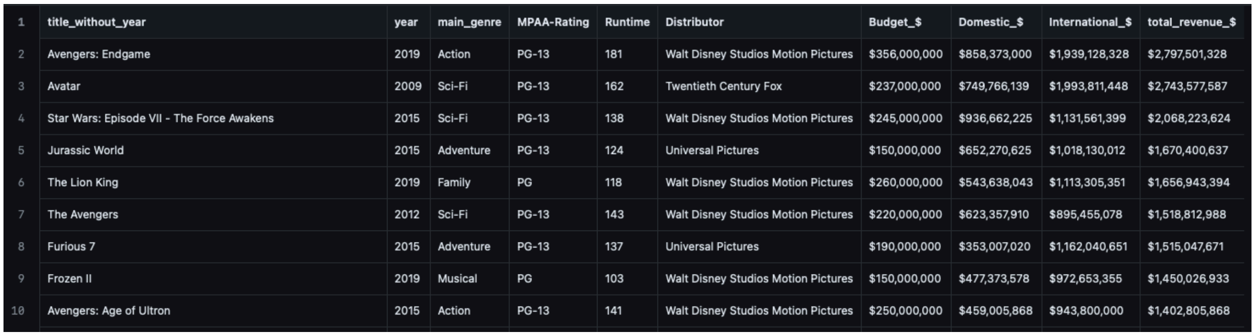
Project 1: Box Office Performance (1999-2019)

For our first group project, we decided to analyze movie box office performance from the year 1999 through 2019. The original dataset taken from Kaggle contained columns for the movie title, domestic gross, international gross, production budget, distributor, MPAA rating, runtime, and genres



Title	Domestic	International	Budget	Distributor	MPAA-Rating	Runtime	Genres
Jurassic World (2015)	652270625	1018130012	150000000.0	Universal Pictures	PG-13	124	Action;Adventure;Sci-Fi
Star Wars: Episode VII - The Force Awakens (2015)	936662225	1131561399	245000000.0	Walt Disney Studios Motion Pictures	PG-13	138	Action;Adventure;Sci-Fi
Avengers: Age of Ultron (2015)	459005868	943800000	250000000.0	Walt Disney Studios Motion Pictures	PG-13	141	Action;Adventure;Sci-Fi
Inside Out (2015)	356461711	501149463	175000000.0	Walt Disney Studios Motion Pictures	PG	95	Adventure;Animation;Comedy;Drama;Family;Fantasy
Furious 7 (2015)	353007020	1162040651	190000000.0	Universal Pictures	PG-13	137	Action;Adventure;Thriller
American Sniper (2014)	350126372	197300000	58800000.0	Warner Bros.	R	133	Action;Biography;Drama;War
Minions (2015)	336045770	823352627	74000000.0	Universal Pictures	PG	91	Adventure;Animation;Comedy;Family;Sci-Fi
The Hunger Games: Mockingjay - Part 2 (2015)	281723902	376620235	160000000.0	Lionsgate	PG-13	137	Action;Adventure;Sci-Fi;Thriller
The Martian (2015)	228433663	401728227	108000000.0	Twentieth Century Fox	PG-13	144	Adventure;Drama;Sci-Fi

To begin the data cleaning, we first had to separate the movie title from the year. The result of this was two separate columns: one with the movie title by itself and one for the year. Another obstacle was the genre column. As you can see, there are multiple genres listed, so we added a script to choose only one genre to make our analysis cleaner. We did this by separating the genres by the semicolon, then we added a randomizer to pick a genre out of the ones that were listed for each movie. This was not the best way to go about this, but it was the most efficient with the time allotted. Next we got rid of the null values in the budget column. As a result, this removed all movies from 2020. There were also a handful of movies scattered throughout the dataset in earlier years that did not contain a budget such as Toy Story 3 (2010). We then created a total revenue column by adding up the domestic and international columns. Finally, we sorted the dataset by the total revenue column in descending order.



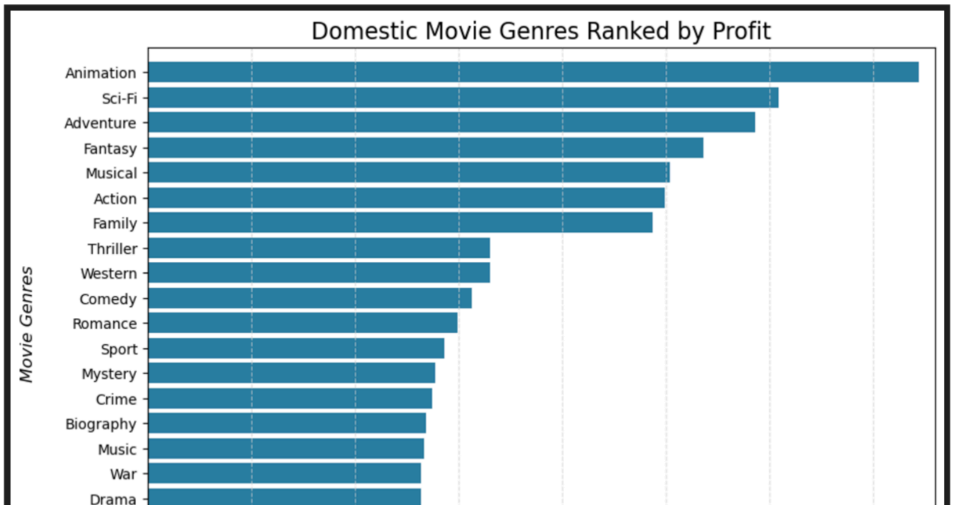
title_without_year	year	main_genre	MPAA-Rating	Runtime	Distributor	Budget_$	Domestic_$	International_$	total_revenue_$
Avengers: Endgame	2019	Action	PG-13	181	Walt Disney Studios Motion Pictures	$356,000,000	$858,373,000	$1,939,128,328	$2,797,501,328
Avatar	2009	Sci-Fi	PG-13	162	Twentieth Century Fox	$237,000,000	$749,766,139	$1,993,811,448	$2,743,577,587
Star Wars: Episode VII - The Force Awakens	2015	Sci-Fi	PG-13	138	Walt Disney Studios Motion Pictures	$245,000,000	$936,662,225	$1,131,561,399	$2,068,223,624
Jurassic World	2015	Adventure	PG-13	124	Universal Pictures	$150,000,000	$652,270,625	$1,018,130,012	$1,670,400,637
The Lion King	2019	Family	PG	118	Walt Disney Studios Motion Pictures	$260,000,000	$543,638,043	$1,113,305,351	$1,656,943,394
The Avengers	2012	Sci-Fi	PG-13	143	Walt Disney Studios Motion Pictures	$220,000,000	$623,357,910	$895,455,078	$1,518,812,988
Furious 7	2015	Adventure	PG-13	137	Universal Pictures	$190,000,000	$353,007,020	$1,162,040,651	$1,515,047,671
Frozen II	2019	Musical	PG	103	Walt Disney Studios Motion Pictures	$150,000,000	$477,373,578	$972,653,355	$1,450,026,933
Avengers: Age of Ultron	2015	Action	PG-13	141	Walt Disney Studios Motion Pictures	$250,000,000	$459,005,868	$943,800,000	$1,402,805,868

Our dataset is now clean and ready to go. So, how do we use this data? We will be analyzing box office statistics and profits, revealing the financial landscape of the film industry, identifying outliers, and assessing the efficacy of high-budget productions.

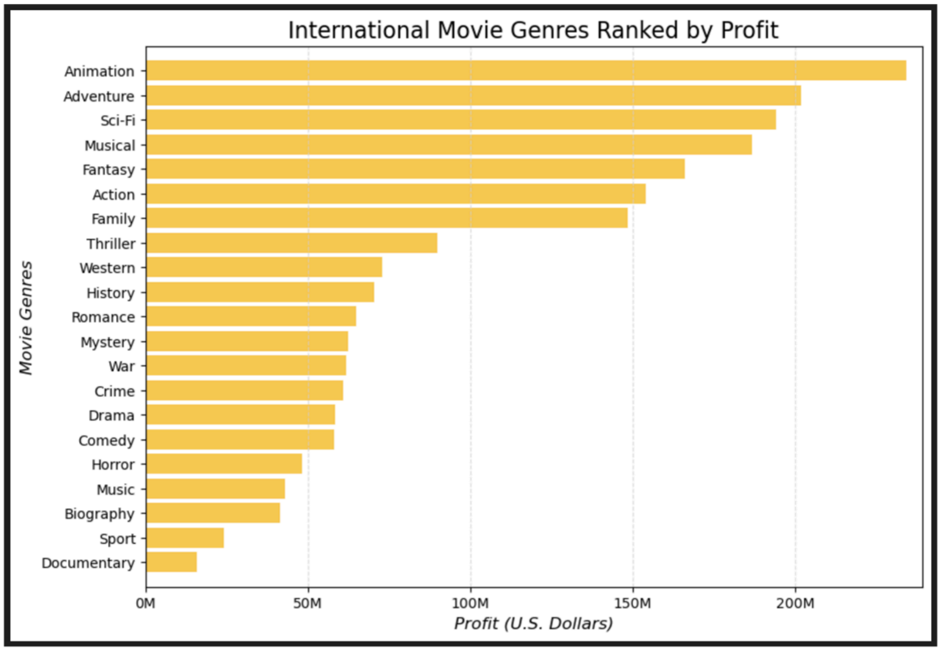
Question 1 (Price) - What are the most popular movie genres? Does genre popularity in the U.S. differ from genre popularity in the international market?

To answer this question, we ranked each of the 21 movie genres by profit. The data were first grouped by genre and then the profits for each genre were summed into two separate categories (domestic and international) to determine how much money each genre made. The data were then sorted from highest to lowest to produce a visual so that each category could be compared.

The data show that of the 21 different movie genres, both the U.S. and international market share the top 9 and the 21st, or 47.62%. The most popular genres across both markets are Animation, Adventure, Sci-Fi, Musical, Fantasy, Action and

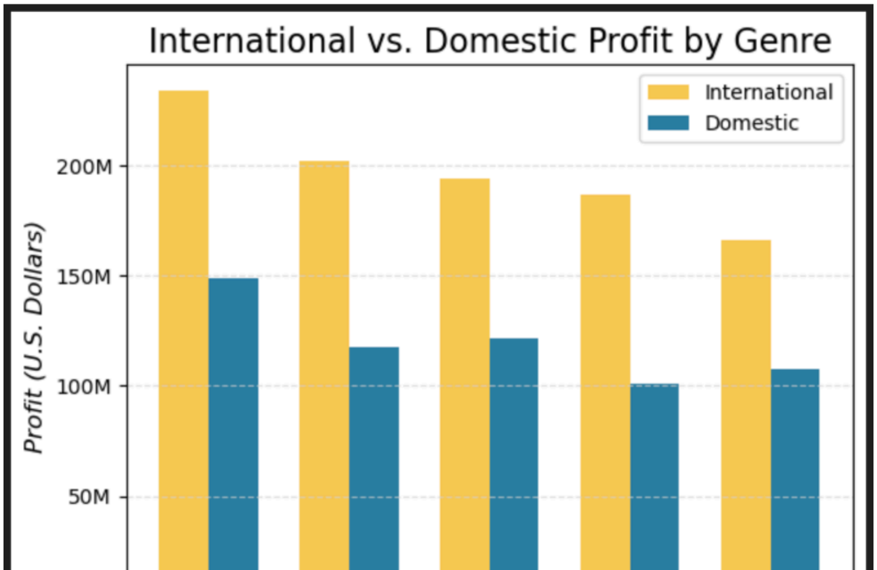


Family. After the top 7, profits decline by approximately $50 million at number 8 (Thriller) and continue to decline steadily; both markets ranking Documentaries last bringing in just $20 million in the U.S. and $15 million abroad.



How do the U.S. profits per genre compare to the international profits for the top 5 genres?

Although the data doesn't specify what countries make up the international movie market, it typically includes the Asia-Pacific region, the Middle East and Africa. The top



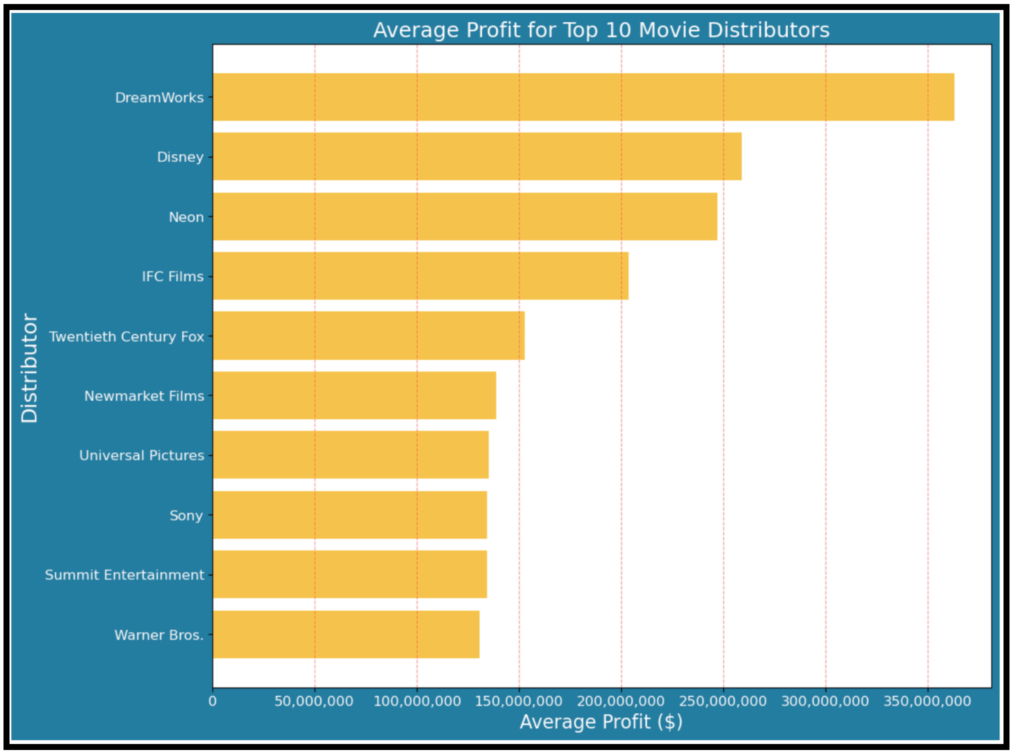
5 genres grossed ~$982 million internationally and ~$596 million domestically. The U.S. market trails behind the international market by just ~$386 million or 39.33%.

Limitations for Question 1

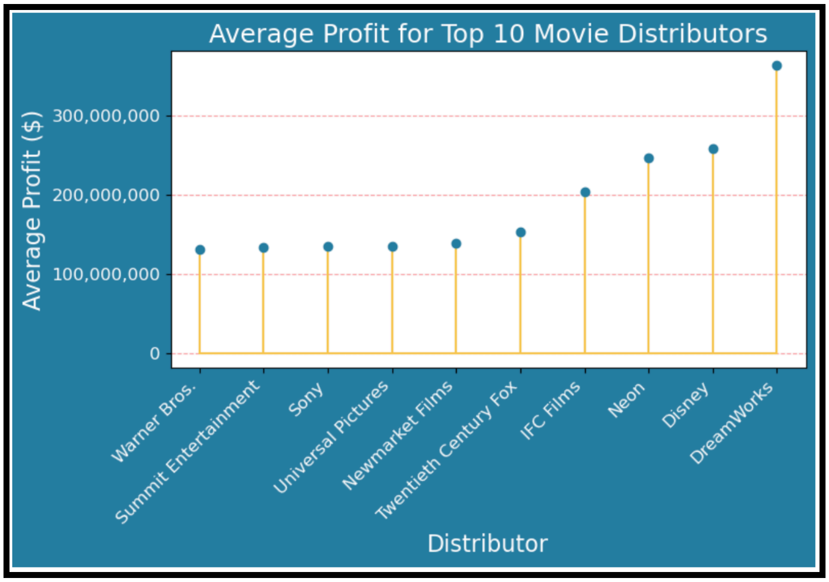
The original data included a list of 4 – 5 movies genres per movie. Avatar, for example, was listed as Action, Adventure, Animated, Sci-Fi, etc. Genres were listed alphabetically, not in order of precedence, so selecting the first genre from the list for each film produced a dataframe full of mostly Action movies. To account for this, we used a randomizer to randomly select one genre from the list for each film. Avatar, therefore, was ultimately listed as a Sci-Fi film because Sci-Fi was the randomly selected genre.

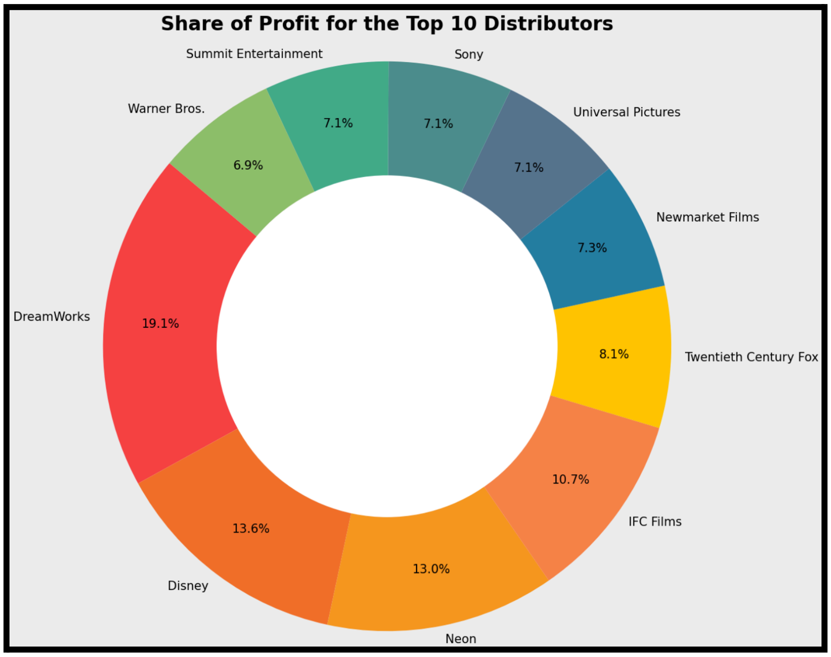
Question 2 (Lee) - Distributor vs Average Profit & Distributor vs Genre

The second research question we hoped to explore was the relationship between our Distributors and two key variables: Average Profit and Genre Distribution. In this pursuit we narrowed our focus to the top 10 highest grossing distributors to see how the leaders of industry behaved. We wanted to see what margins our industry leaders were leading by in regards to profit. With genre count, we hoped to see what the distribution looked like across our distributors. Did they produce similar quantities across all genres and if not maybe that could explain what set certain distributors apart from their peers. With this in mind let’s see what our visualizations tell us:

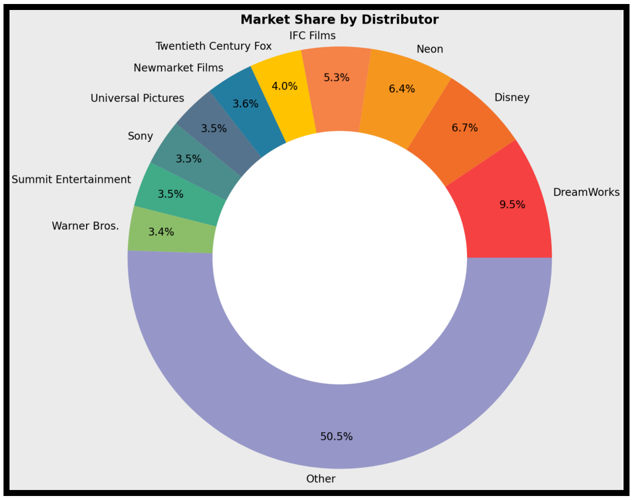


The chart above shows that our top 10 distributors are making about $125,000,000 per movie. We also see that DreamWorks leads the pack by a significant margin. This wasn’t the case when I first produced this graph but I noticed something that was worth further cleaning: DreamWorks and DreamWorks Distribution were two separate data entries. Another note worth making is that our dataset ends in 2019 and Disney has since acquired 20th Century Fox which would put them up there with Dreamworks current day. Perhaps a future project could help us understand why these distributors are on top by analyzing which franchises perform the best for each. Another way to visualize this could be a lollipop chart or a pie chart as follows:

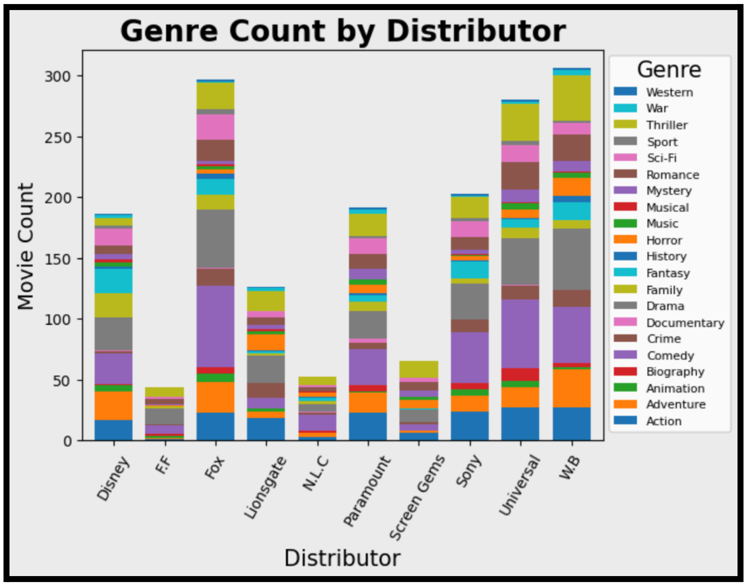




In case we wanted that greater market context I created a second version of this pie chart that puts that into perspective:



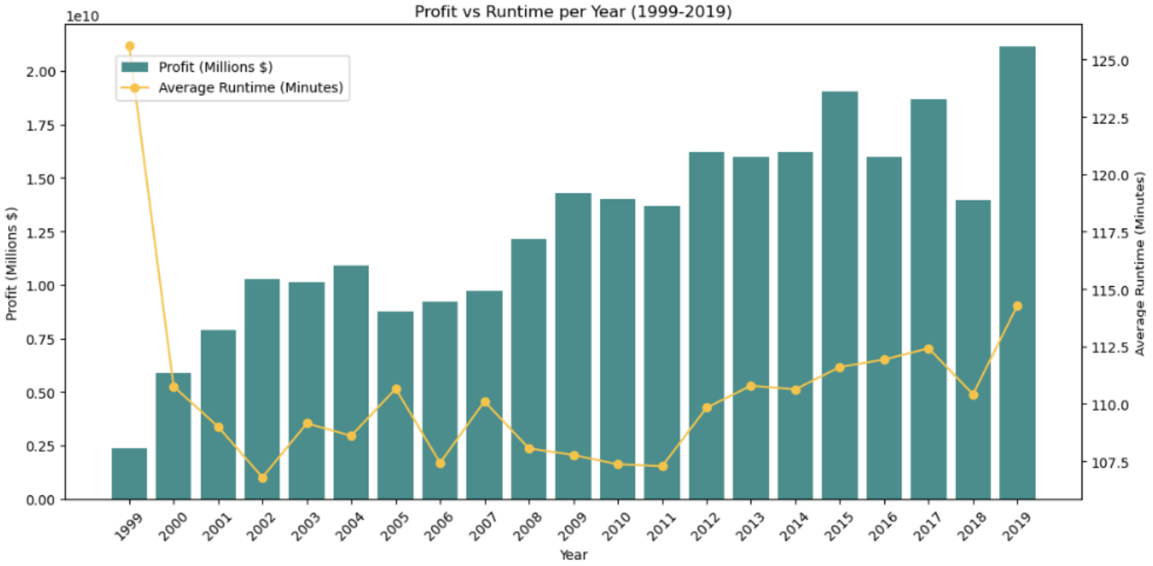
The second part of this research question was to examine the genre count. We can see by the following stacked bar chart that each distributor has a similar distribution of genres in their portfolios and we can see that Thriller, Drama, Action, Comedy, and Adventure emerge as our most popular categories. This is of course heavily limited by the way in which we cleaned our genres in the data cleaning process.



Question 3 (Basazinew)

We aimed to explore the relationship between movie runtimes and their profitability. To start, I used a groupby operation to calculate the average runtime for each year, providing an overall view of which years had the longest movies on average. This preliminary step set the stage for further analysis to determine if those years also experienced higher profitability.

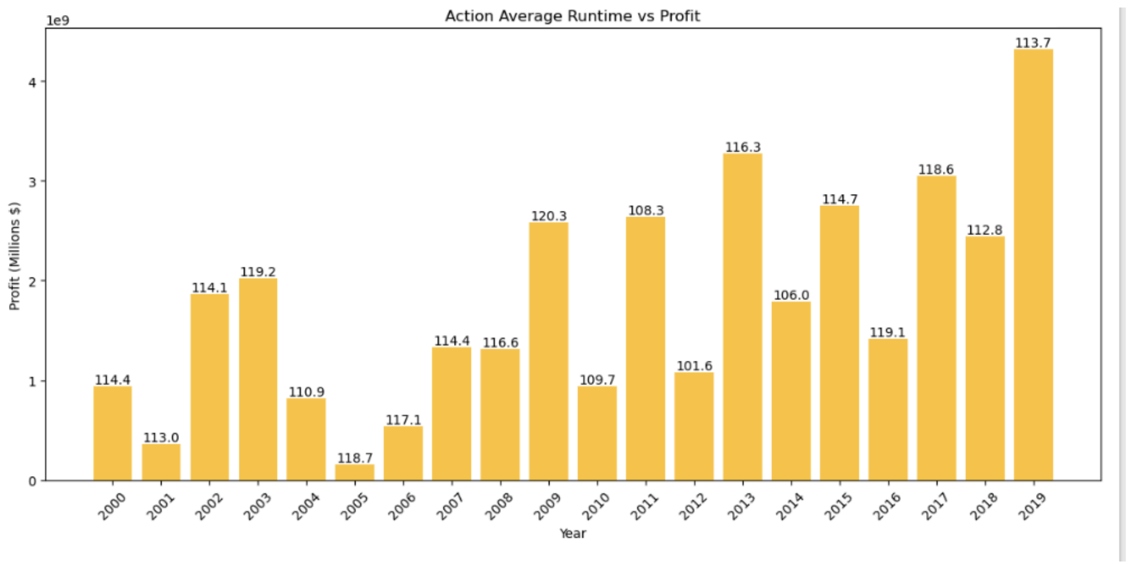
I filtered the data for the years 1999 to 2019 and grouped it by year. Then, I used an aggregate function to calculate the average runtime and profit for each year. I extracted the years, average runtimes, and profits, and created a bar chart to visually compare profit and runtime from 1999 to 2019. This first bar chart provides a visual representation to determine if there is a correlation between the year with the longest average runtime and its corresponding profit, exploring if a longer runtime positively influences profitability.



Following this, I created a scatter plot and calculated the correlation coefficient, which is 0.28. This reveals a weak positive linear relationship between runtime and profit. Although this suggests that longer runtimes are not strong predictors of higher profits, it doesn't completely disprove a correlation. Therefore, further investigation is warranted.

Next, I explored whether genres with longer runtimes are more profitable. I created a bar graph showing each genre's average runtime alongside its average profit, which illustrated the relationship between length and profitability across different genres. To do this, I filtered the original DataFrame df to create df_genre, containing only rows where main_genre matches the specified genre. Using a groupby operation on the year column, I aggregated the data to find the average runtime and total profit for each year, summarizing these metrics by genre and year. I extracted year, Runtime, and profit from the genre_stats DataFrame into years, average_runtimes, and profits_millions, preparing the data for plotting or further analysis by making these variables readily accessible. I then created a bar chart of year and profit using an f-string with {specific_genre}, allowing for easy generation of bar charts for any chosen genre.

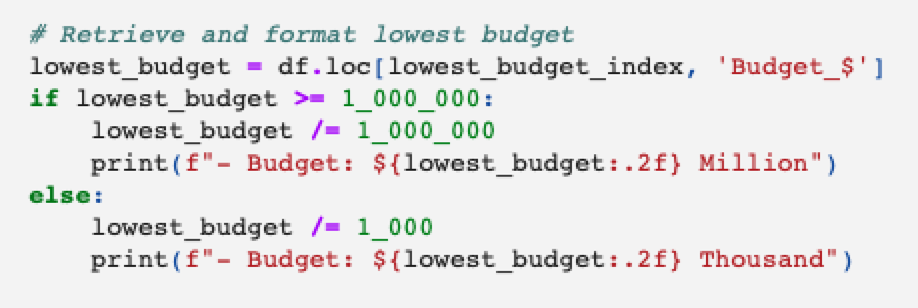
The first bar chart displayed is generated by calling profit_vs_runtime_genre('Action'). This function serves as a utility for analyzing and visualizing the financial and runtime data of movies within a specific genre. For example, calling profit_vs_runtime_genre('Action') analyzes and displays this data specifically for Action movies.



Finally, I presented several more bar graphs to further demonstrate that there is no correlation between longer movie runtimes and their profit across various genres. It's worth noting that in the last five examples, some bars appear empty due to the randomizer applied during our data cleaning.

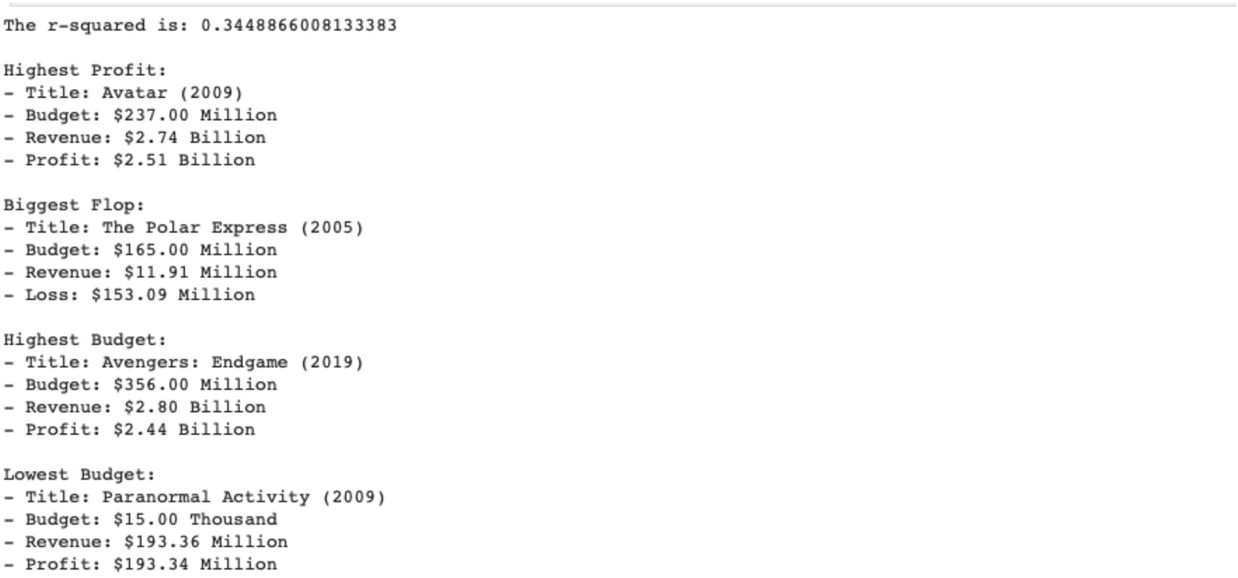
Regression (Vick)

For our regression model, we looked at the correlation between a movie's production budget and its profit. In the dataset we cleaned up together, we neglected to create a profit column. So, the first thing I needed to do was write a quick script to create a profit column. I simply subtracted the value in the budget column from the value in the total revenue column. Voila, we now have a profit column to work with. The next challenge is to define the linear regression plot: slope, intercept, r_value, p_value, std_err **=** linregress(df[x_col], df[y_col]). Next, I created a scatter plot to visualize the linear regression, added the appropriate labels, and changed the colors and size for the data points as well as the linear regression line. In order to come to a conclusion, I added code to print the r-value as well as movies with the highest profit, highest loss, highest budget, and lowest budget. Honestly, The supplementary code to print out all the relevant information was the longest part of this coding session. For each of the results, I wanted the movie's title, year, budget, revenue, and profit printed neatly and formatted correctly. I had to write statements so that if the budget was in the millions, then it would be presented as millions and not billions and so on and so forth. However, the lowest budget movie was in the thousands, so I had to write a script so that the budget for Paranormal Activity showed up as thousands instead of millions. I kept running into errors where the value would be represented as a 0. The value was too small to be accurately represented. This part gave me the most grief, but I was finally able to get the accurate budget for the movie.

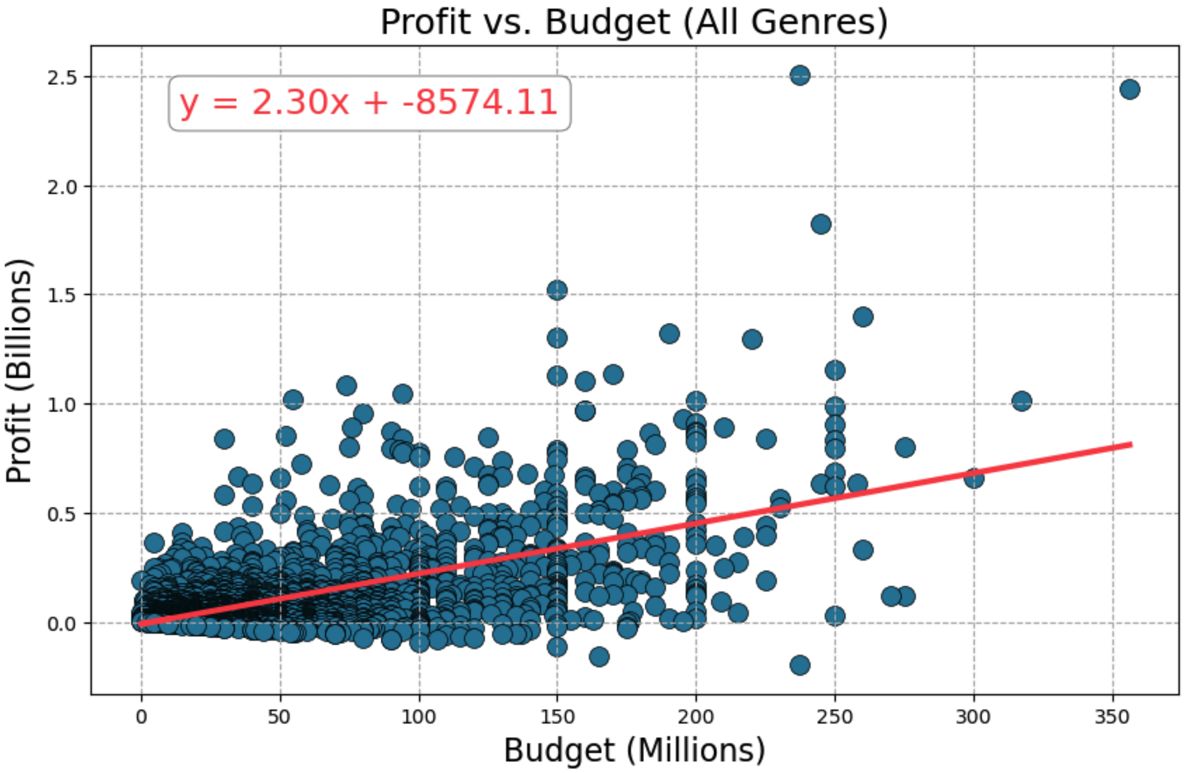


```
# Retrieve and format lowest budget
lowest_budget = df.loc[lowest_budget_index, 'Budget_$']
if lowest_budget >= 1_000_000:
    lowest_budget /= 1_000_000
    print(f"- Budget: ${lowest_budget:.2f} Million")
else:
    lowest_budget /= 1_000
    print(f"- Budget: ${lowest_budget:.2f} Thousand")
```

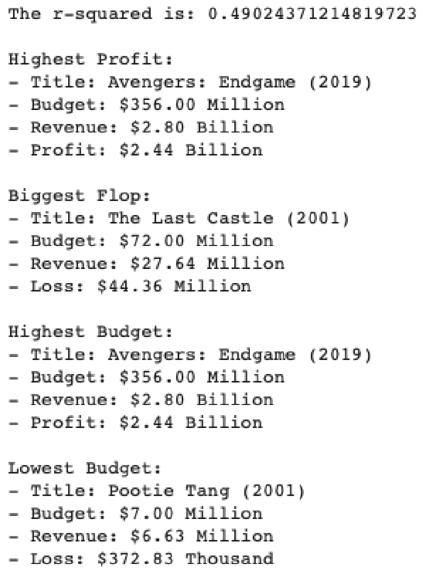
Finally, the coding is done and we can show our linear regression plot and print all the necessary data to come to a conclusion. When I ran the linear regression code for all genres, I got an r-value of 0.34. This means that a movie's production budget accounts for roughly 34% of its success regardless of what genre it is.



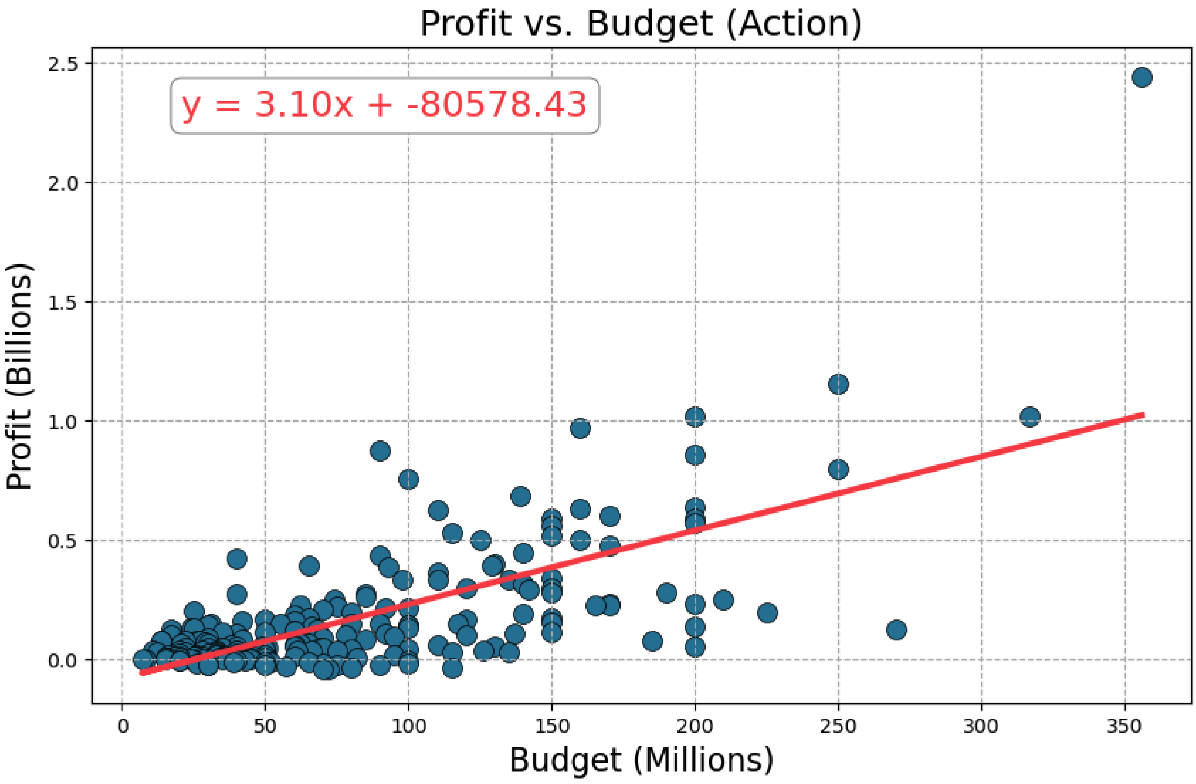
```
The r-squared is: 0.3448866008133383

Highest Profit:
- Title: Avatar (2009)
- Budget: $237.00 Million
- Revenue: $2.74 Billion
- Profit: $2.51 Billion

Biggest Flop:
- Title: The Polar Express (2005)
- Budget: $165.00 Million
- Revenue: $11.91 Million
- Loss: $153.09 Million

Highest Budget:
- Title: Avengers: Endgame (2019)
- Budget: $356.00 Million
- Revenue: $2.80 Billion
- Profit: $2.44 Billion

Lowest Budget:
- Title: Paranormal Activity (2009)
- Budget: $15.00 Thousand
- Revenue: $193.36 Million
- Profit: $193.34 Million
```



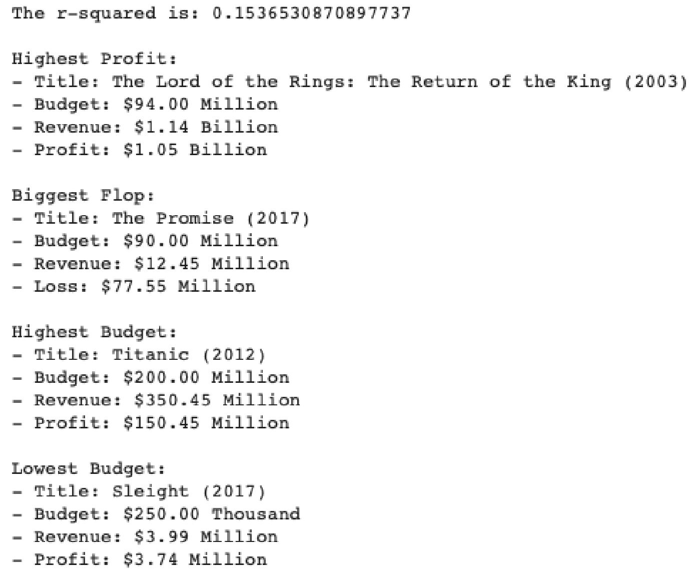
Next, I applied the same linear regression model to individual genres to find any outliers. Action came back with a whopping r-value of 0.49. According to the data, an action movie’s success is heavily dependent on its budget. This makes sense because action movies require larger budgets for all the explosions, CGI, and massive sets to work on. So, it stands to reason that the bigger the explosion then the higher the profits will be. A sentiment that Michael Bay takes to heart.



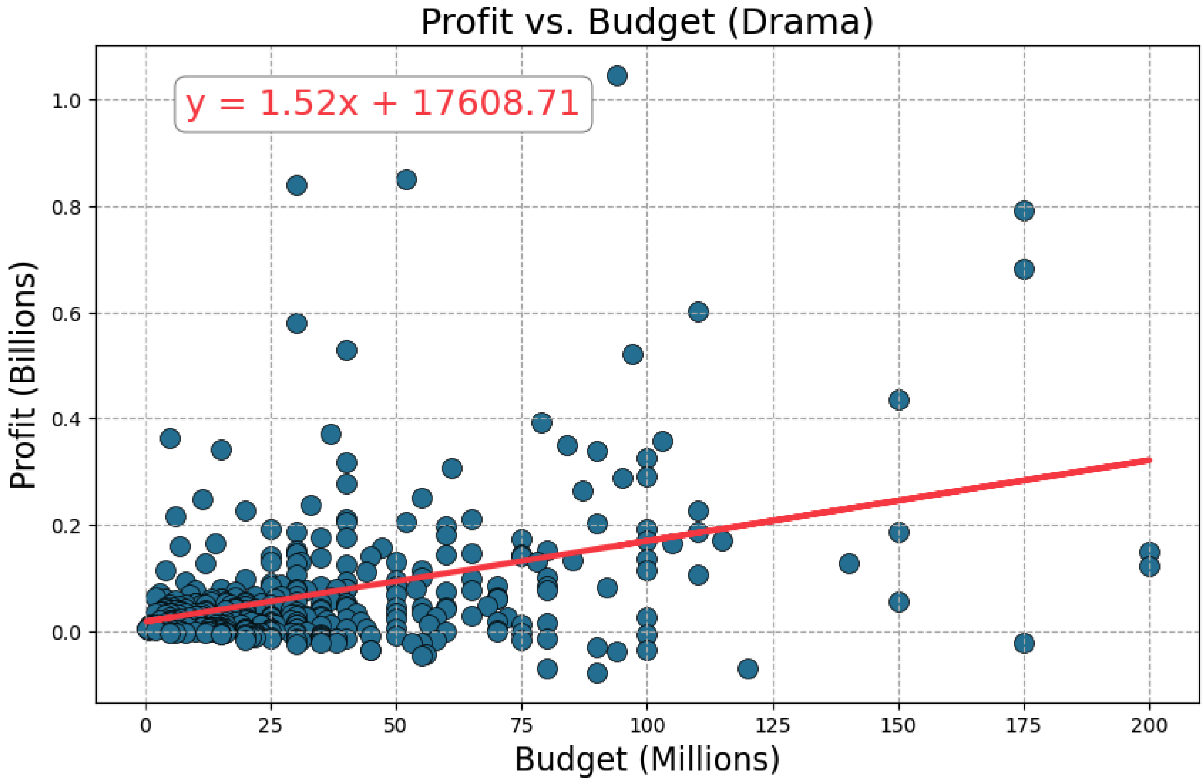
```
The r-squared is: 0.49024371214819723

Highest Profit:
- Title: Avengers: Endgame (2019)
- Budget: $356.00 Million
- Revenue: $2.80 Billion
- Profit: $2.44 Billion

Biggest Flop:
- Title: The Last Castle (2001)
- Budget: $72.00 Million
- Revenue: $27.64 Million
- Loss: $44.36 Million

Highest Budget:
- Title: Avengers: Endgame (2019)
- Budget: $356.00 Million
- Revenue: $2.80 Billion
- Profit: $2.44 Billion

Lowest Budget:
- Title: Pootie Tang (2001)
- Budget: $7.00 Million
- Revenue: $6.63 Million
- Loss: $372.83 Thousand
```



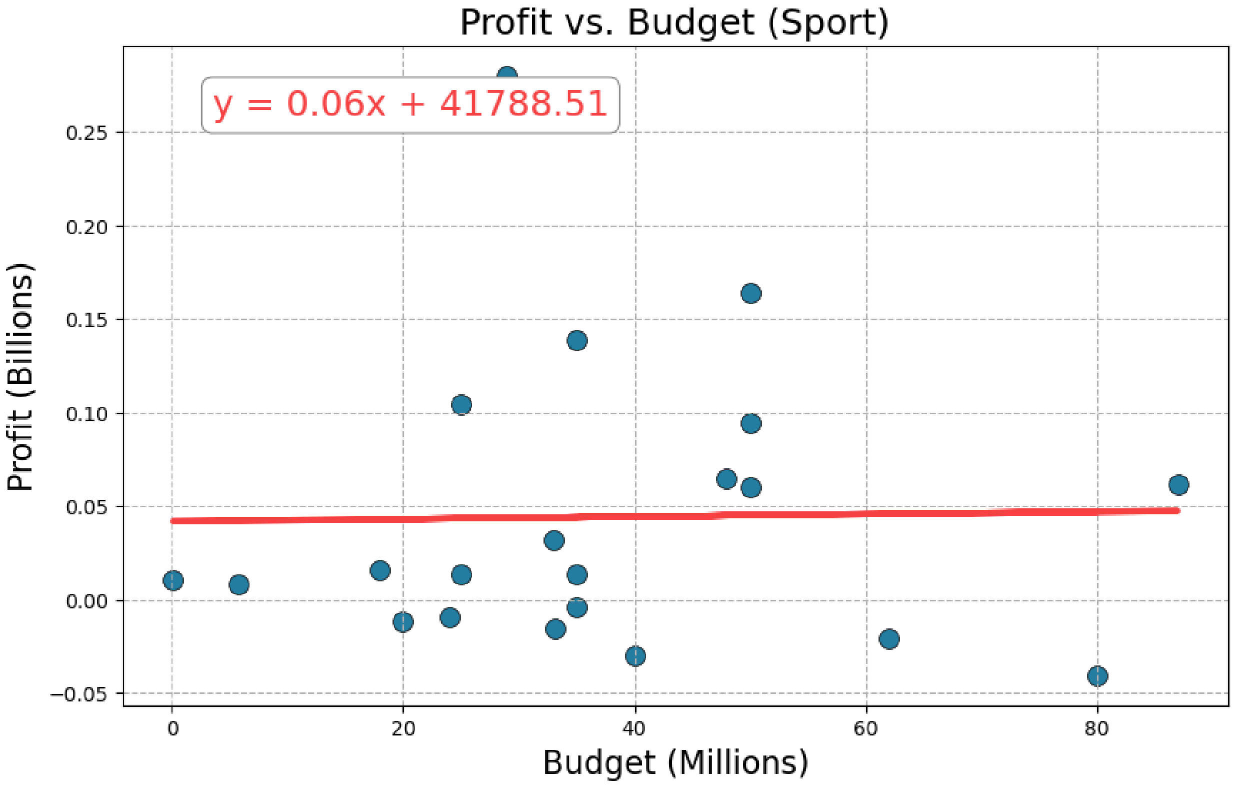
Next up on the chopping block was drama. Drama returned an r-value of 0.15, so their profits aren't as dependent on the budget as action movies are. Dramas typically don’t need all the big explosions and CGI aliens that action movies do. Dramas depend more on good writing and the emotions being conveyed on the screen. You don’t typically need a laser beam firing up in the sky to make someone cry.



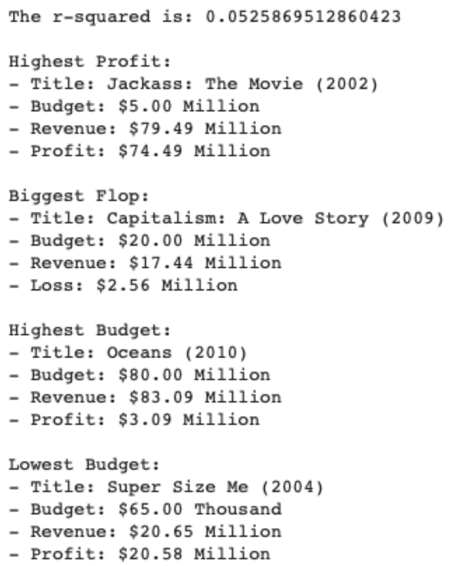
```
The r-squared is: 0.1536530870897737

Highest Profit:
- Title: The Lord of the Rings: The Return of the King (2003)
- Budget: $94.00 Million
- Revenue: $1.14 Billion
- Profit: $1.05 Billion

Biggest Flop:
- Title: The Promise (2017)
- Budget: $90.00 Million
- Revenue: $12.45 Million
- Loss: $77.55 Million

Highest Budget:
- Title: Titanic (2012)
- Budget: $200.00 Million
- Revenue: $350.45 Million
- Profit: $150.45 Million

Lowest Budget:
- Title: Sleight (2017)
- Budget: $250.00 Thousand
- Revenue: $3.99 Million
- Profit: $3.74 Million
```



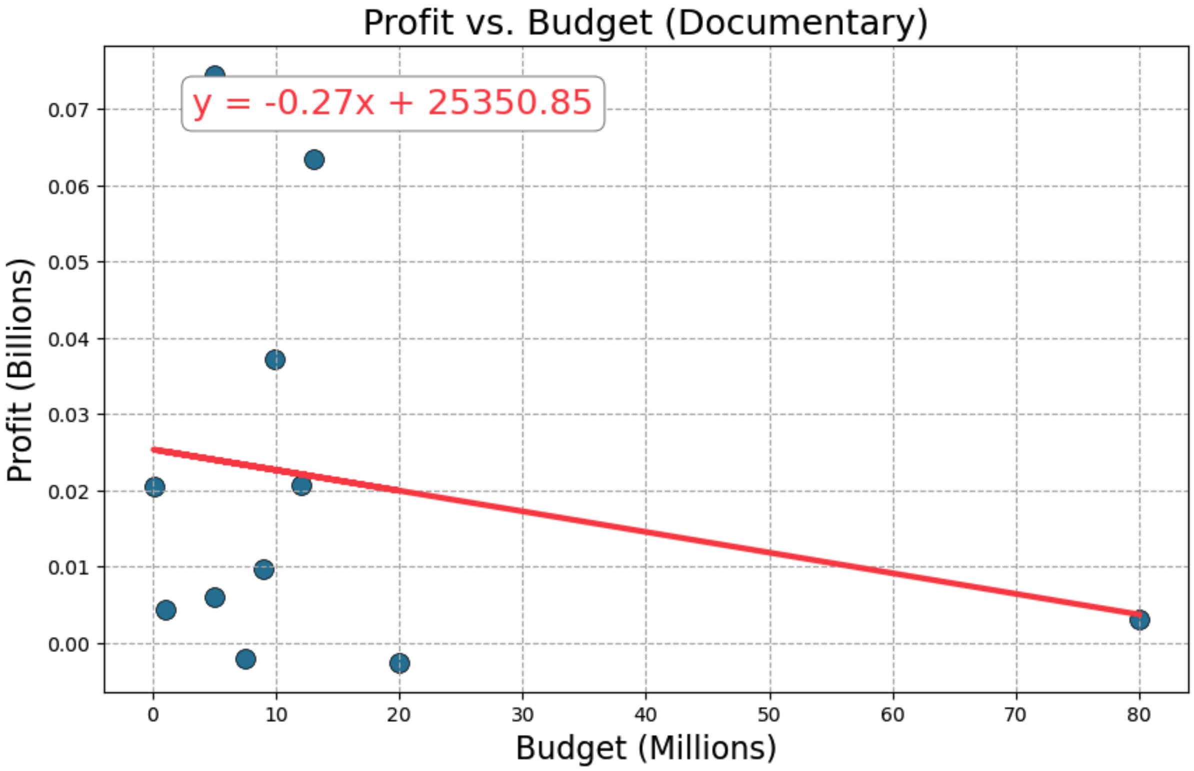
A genre with virtually no correlation between its budget and profitability was sports (sorry professor). The r-value for sports movies was 0.0003. Honestly, I was a bit surprised, but I guess when it comes to sports movies it's more about whether the topic was a beloved enough moment in sports history. Sports movies were my biggest surprise for sure, but at least it wasn’t entirely negative.



The only negative correlation I found was for documentaries. The r-value for that genre came out as 0.05. This one should be obvious to any casual viewer. Documentaries are mostly information based consisting of archived media and brightly lit interviews. They do not require higher budgets whatsoever. The exception may be nature documentaries, but even then they do not require a budget in the tens of millions. The highest budget documentary in the dataset was for Oceans (2010) for $80 million and only returned a profit of $3.09 million. You have to ask yourself if that is really worth it from a business perspective.



```
The r-squared is: 0.0525869512860423

Highest Profit:
- Title: Jackass: The Movie (2002)
- Budget: $5.00 Million
- Revenue: $79.49 Million
- Profit: $74.49 Million

Biggest Flop:
- Title: Capitalism: A Love Story (2009)
- Budget: $20.00 Million
- Revenue: $17.44 Million
- Loss: $2.56 Million

Highest Budget:
- Title: Oceans (2010)
- Budget: $80.00 Million
- Revenue: $83.09 Million
- Profit: $3.09 Million

Lowest Budget:
- Title: Super Size Me (2004)
- Budget: $65.00 Thousand
- Revenue: $20.65 Million
- Profit: $20.58 Million
```



Call to Action (Vick)

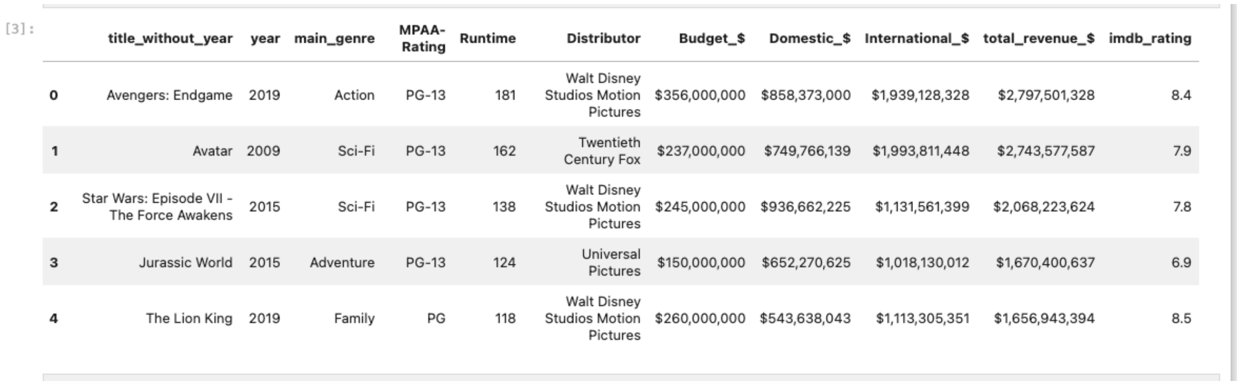
Our call to action can be directed at three separate groups. For the filmmakers and studios, they should embrace calculated risks. While high budget films can bring in big box office numbers, our analysis suggests that lower-budget films can also be quite profitable. For the investors, they should consider a diverse portfolio of different genres with varying budgets to mitigate risk. Lastly, general audiences should watch movies across different genres and budgets to foster a vibrant and diverse film industry. We don't want the entire industry to be solely Marvel movies do we?

Bias & Limitations (Vick)

While the dataset originally had nearly 3,000 rows of data, it didn't come without its limitations. Firstly, the information was not complete. We were missing budgets from nearly 1,000 movies. Second, the data only went up to 2020 at the latest and we were only able to analyze the data up to 2019. Well, a lot has happened since 2019. The movie theater business has been flipped upside down and streaming from the comfort of your own home has taken over. We would have to completely reevaluate our conclusions if we were to include everything that has happened since COVID-19. Furthermore, our dataset did not include any ratings from IMDb or review aggregators such as Rotten Tomatoes or Metacritic. We can hypothesize that both critical and audience reception plays a massive impact on a movie's profitability, but we did not have that information readily available. Lastly, a limitation that we placed on ourselves was the genre randomizer. We did not always get the most appropriate genre assigned to each movie. For example, I would not consider the movie Pootie Tang to be an action movie, but the randomizer chose action out of the other possible genres listed for Pootie Tang. Hilariously, Jackass was considered a documentary. While the footage in Jackass is real, I would not consider documentary to be the most appropriate genre. Most of us would probably consider it to be a comedy. So, any analysis based on genre is automatically negligible due to the lack of sophistication of the randomizer.

Future Work (Vick)

Based directly on the limitations of the project, we could easily update it for any future work. We can update the dataset with movies that were released after COVID-19. We can also add ratings to all the movies currently in the dataset. I actually created a script that used an API key that added IMDb ratings to all the movies. Unfortunately, I did not think about it until the Thursday before the project was due. We were told there wouldn't be enough time and to ignore it for the work we already had. With all these updates to the dataset, we can portray a more accurate representation of the movie industry today and where it is headed.



[3]:

	title_without_year	year	main_genre	MPAA-Rating	Runtime	Distributor	Budget_$	Domestic_$	International_$	total_revenue_$	imdb_rating
0	Avengers: Endgame	2019	Action	PG-13	181	Walt Disney Studios Motion Pictures	$356,000,000	$858,373,000	$1,939,128,328	$2,797,501,328	8.4
1	Avatar	2009	Sci-Fi	PG-13	162	Twentieth Century Fox	$237,000,000	$749,766,139	$1,993,811,448	$2,743,577,587	7.9
2	Star Wars: Episode VII - The Force Awakens	2015	Sci-Fi	PG-13	138	Walt Disney Studios Motion Pictures	$245,000,000	$936,662,225	$1,131,561,399	$2,068,223,624	7.8
3	Jurassic World	2015	Adventure	PG-13	124	Universal Pictures	$150,000,000	$652,270,625	$1,018,130,012	$1,670,400,637	6.9
4	The Lion King	2019	Family	PG	118	Walt Disney Studios Motion Pictures	$260,000,000	$543,638,043	$1,113,305,351	$1,656,943,394	8.5

Works Cited

Fritz, Brian. *OMDb*, 17 June 2024, www.omdbapi.com.

Larue, Luke. "Movie Attributes for 3400 Movies from 2000-2020." *Kaggle*, 17 June 2024, www.kaggle.com/datasets/lukelarue/movie-attributes-for-3400-movies-from-2000 2020.

Box Office Mojo, 17 June 2024, www.boxofficemojo.com.