**Project Overview**

Our hypothetical Company has announced that they would like to get into the movie industry. They will be creating a studio, however they have no knowledge of the movie industry. My goal is to collect, clean, and analyze movie data from a variety of sources so that I can provide recommendations to the company that will allow them to be successful in the movie industry.

**Python Libaries Used**

***pandas, numpy, seaborn, matplotlip, datetime***

**Data and Exploration**

**Web-scrapping**

The web-scraped data used in this project was collected from the following sources:

1. <https://www.the-numbers.com/movie/budgets/all/1>
2. <https://www.imdb.com/search/title/?title_type=feature&num_votes=5000,&languages=en&sort=boxoffice_gross_us,desc&start=1&explore=genres&ref_=adv_nx>
3. <https://www.boxofficemojo.com/chart/most_theaters/?by_studio_type=major>
4. <https://www.moviefone.com/movies/2019/?page=1>
5. <https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films>

**Exploration Questions**

In my analysis I explore and answer the following questions:

1. What are the most profitable movies and how much should you spend?
2. Which movie genres are most commonly produced and does quantity equate to higher net profits?
3. What is the best time of the year to release a movie?
4. Which actors and directors tend to add the most value?
5. How much money should you spend to win an Oscar?
6. What impact, if any, does runtime and movie rating have on Net Profit, Profit Margin and IMDb rating?
7. Sticking to our analysis of Net Profit and Profit Margin, what should our Company determine to be the baseline for sustainable success?
8. Based on the success of current competitors, which should we look to for best practices?

**Question 1: What are the most profitable movies and how much should you spend?**

To answer this question and provide a recommendation we'll make use of a budgets dataframe called imdb\_budgets\_df. Our analysis will require that we use the data to calculate profit and profit margin.

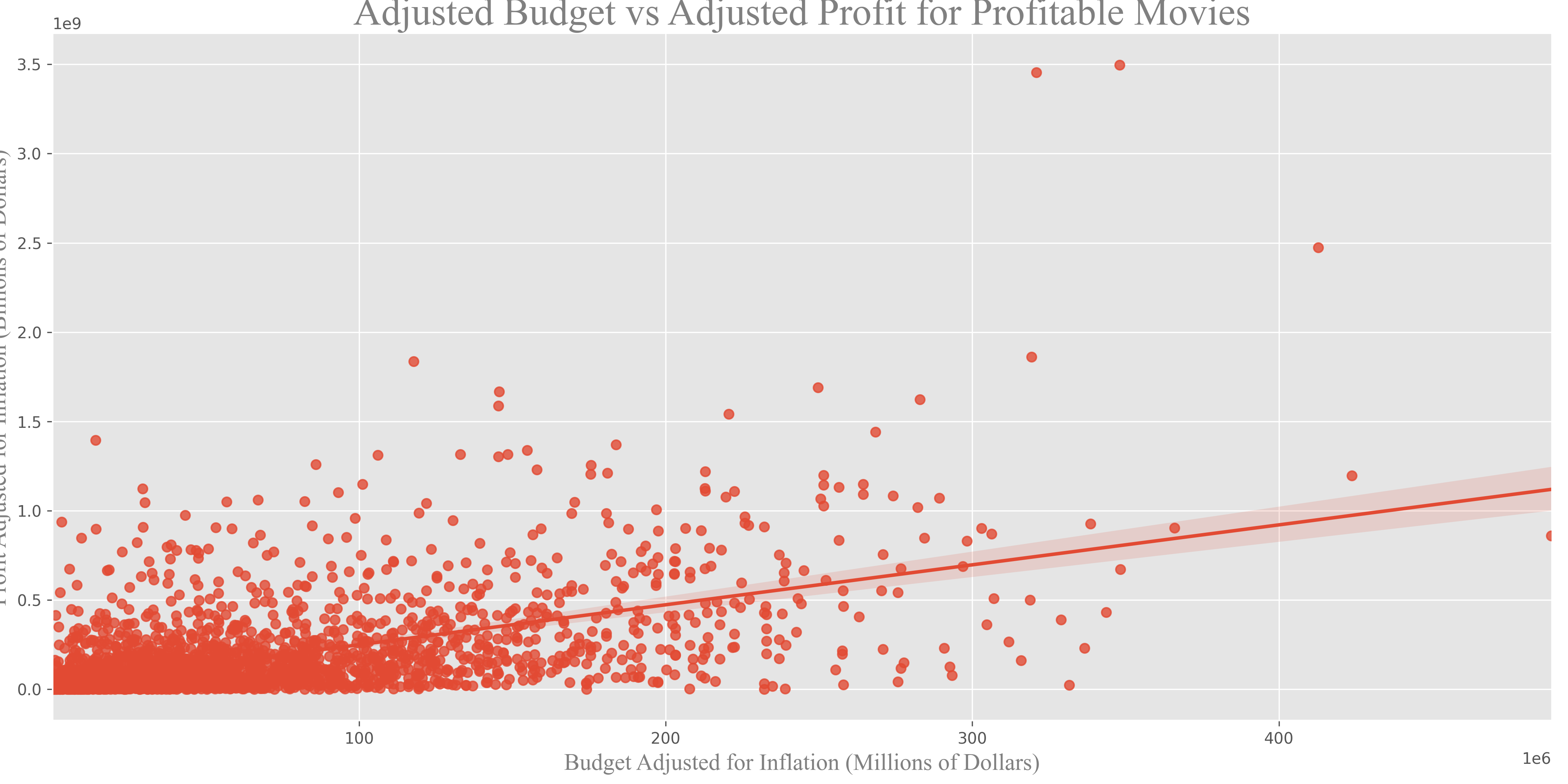
imdb\_budgets\_df['Profit'] = imdb\_budgets\_df['Worldwide Gross'] - imdb\_budgets\_df['Production Budget']

imdb\_budgets\_df['Profit\_Margin'] = (imdb\_budgets\_df['Worldwide Gross'] -

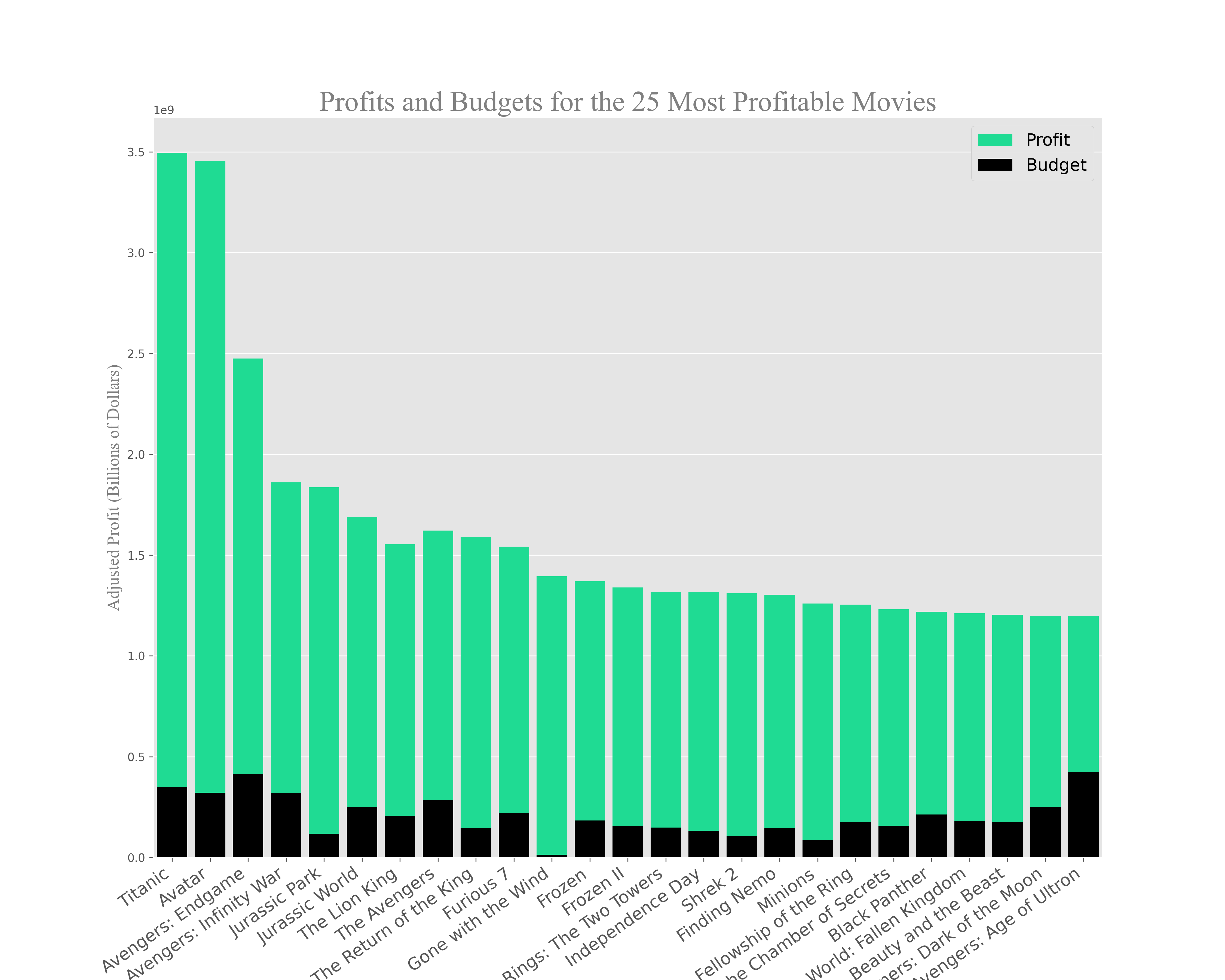
imdb\_budgets\_df['Production Budget'])/imdb\_budgets\_df['Worldwide Gross']

We will also create two columns called Adjusted\_Budget and Adjusted\_Profit where we recalculate a movie's budget and profit to account for inflation and allow us to perform analysis based on the value of the 2020 dollar.

We examine the overall trend of budget versus profit to see if there's any correlation.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/BudgetVProfit.png)

We also take a look at the top 25 movies in terms of profit to understand their financial success and how closely we should attempt to emulate their budget. We can see in the graph below there are a few outliers so the median will end up being more useful in determining our final budget.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/ProfitBudgetTop25.png)

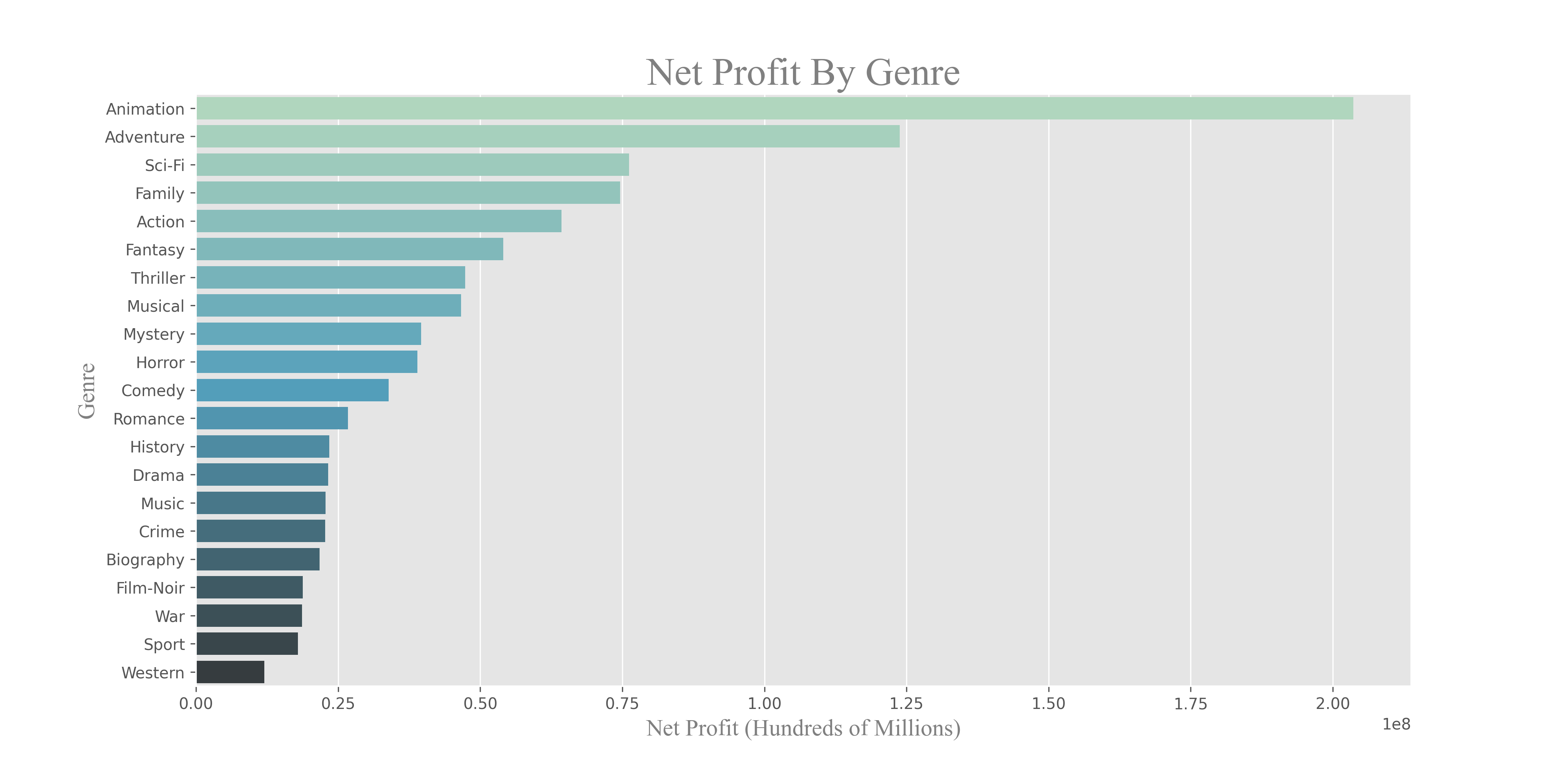
**Question 1 Conclusion:** Our Company should budget $82,250,000 for a movie and that budget should correlate with a profit margin of 80%.

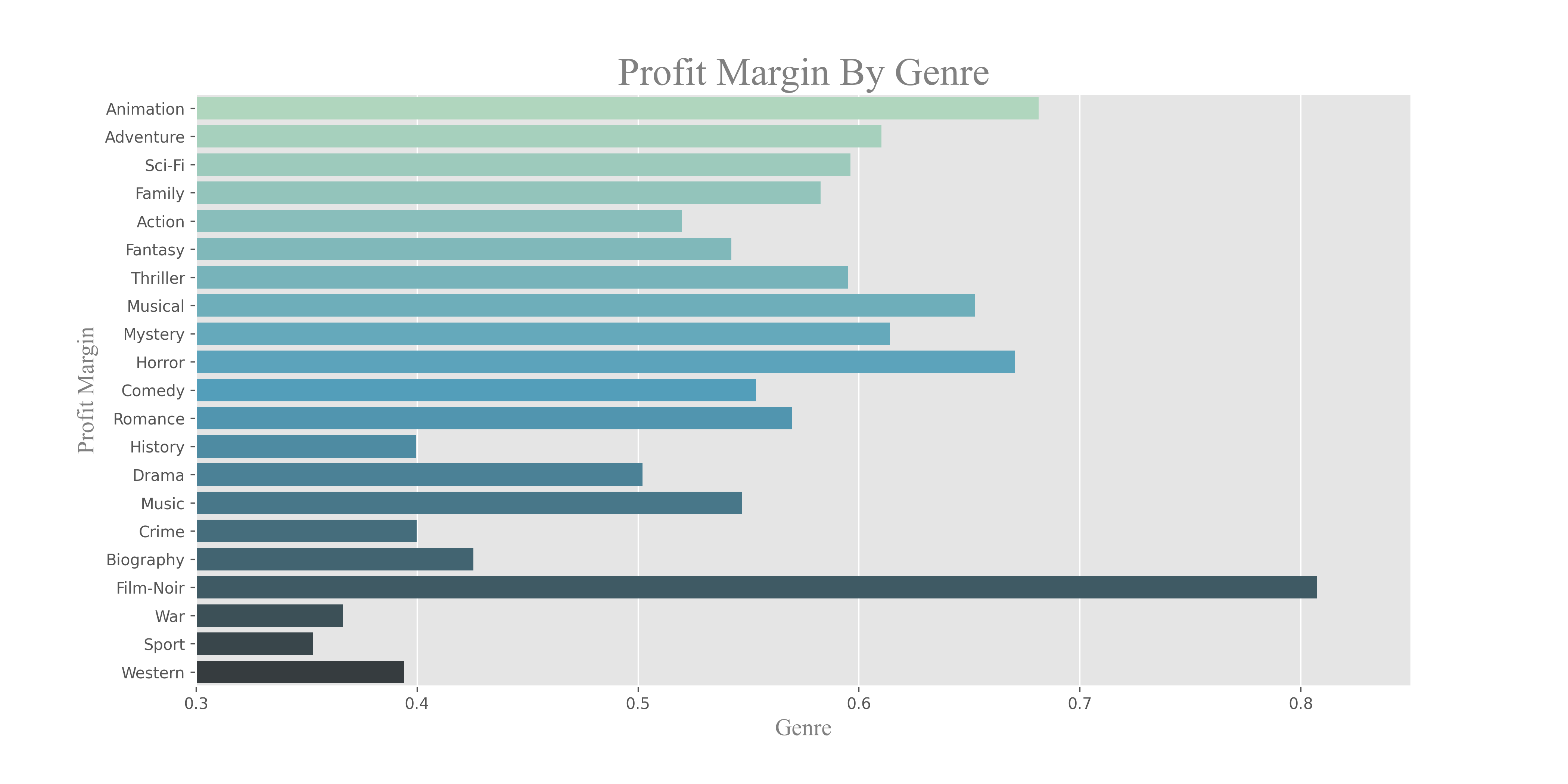
**Question 2: Which movie genres are most commonly produced and does quantity equate to higher net profits?**

We first count the number of movies in each genre and plot those results on a bar graph.

m\_by\_genre = genre\_budgets\_df.groupby('Genre', as\_index=False)['Movie'].count().sort\_values(by='Movie', ascending=False)

Using the same groupby method, we select the median net profit and profit margin for each genre. We use the median in this case as the mean is likely skewed by outliers. Outliers could either be movies with enormous profits or movies having negative profit.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/NetProfitGenre.png)

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/ProfitMarginGenre.png)

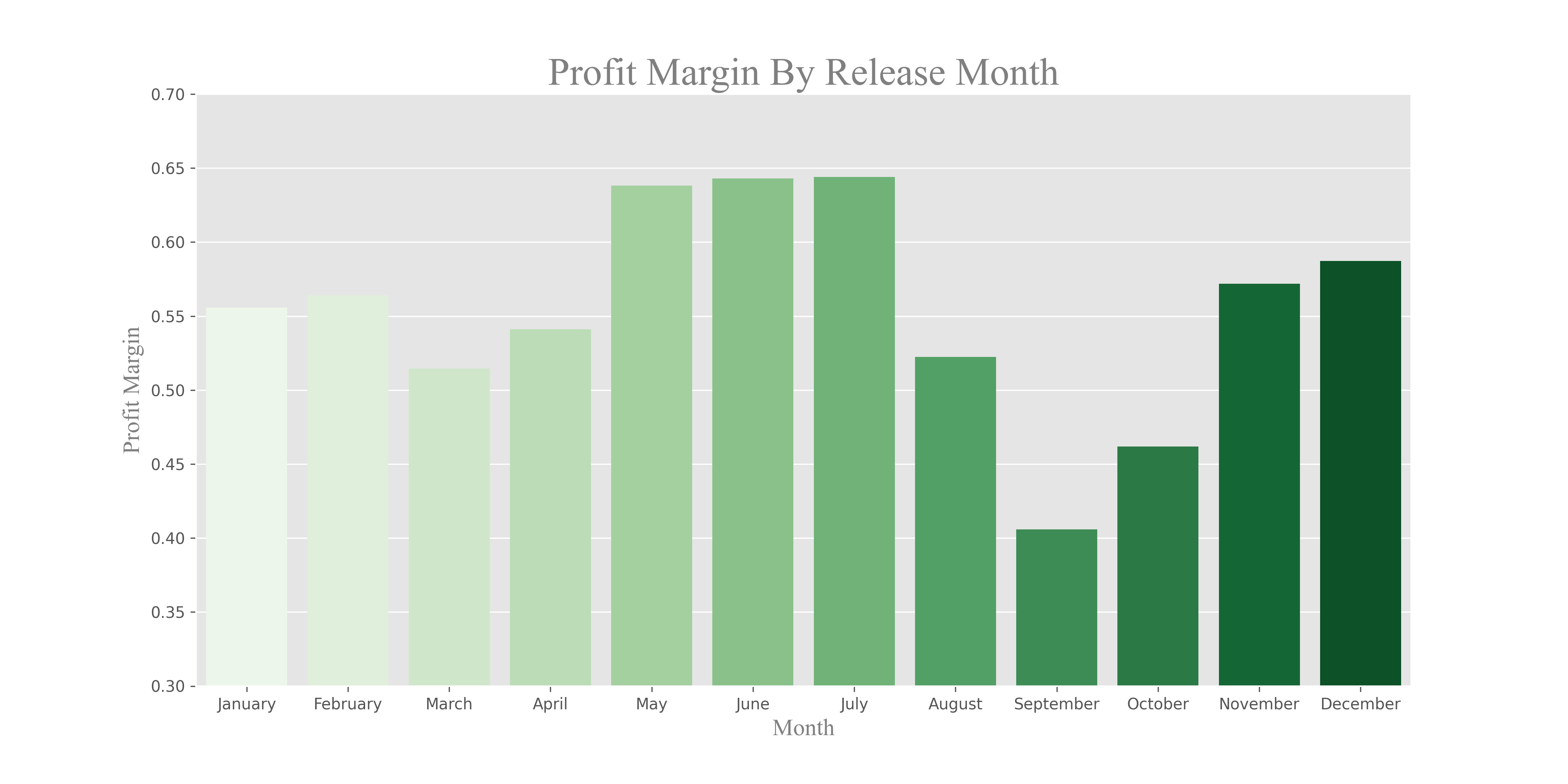
Lastly, we look at the percent of net profit by genre. This informs us as to how Our Company should allocate their movie budget to various films.

**Question 2 Conclusion:** Animation, adventure, and sci-Fi have the highest net profit of all genres. Analysis of profit margin shows that in addition animation, adventure, and sci-fi, horrors and musicals also have financial success.

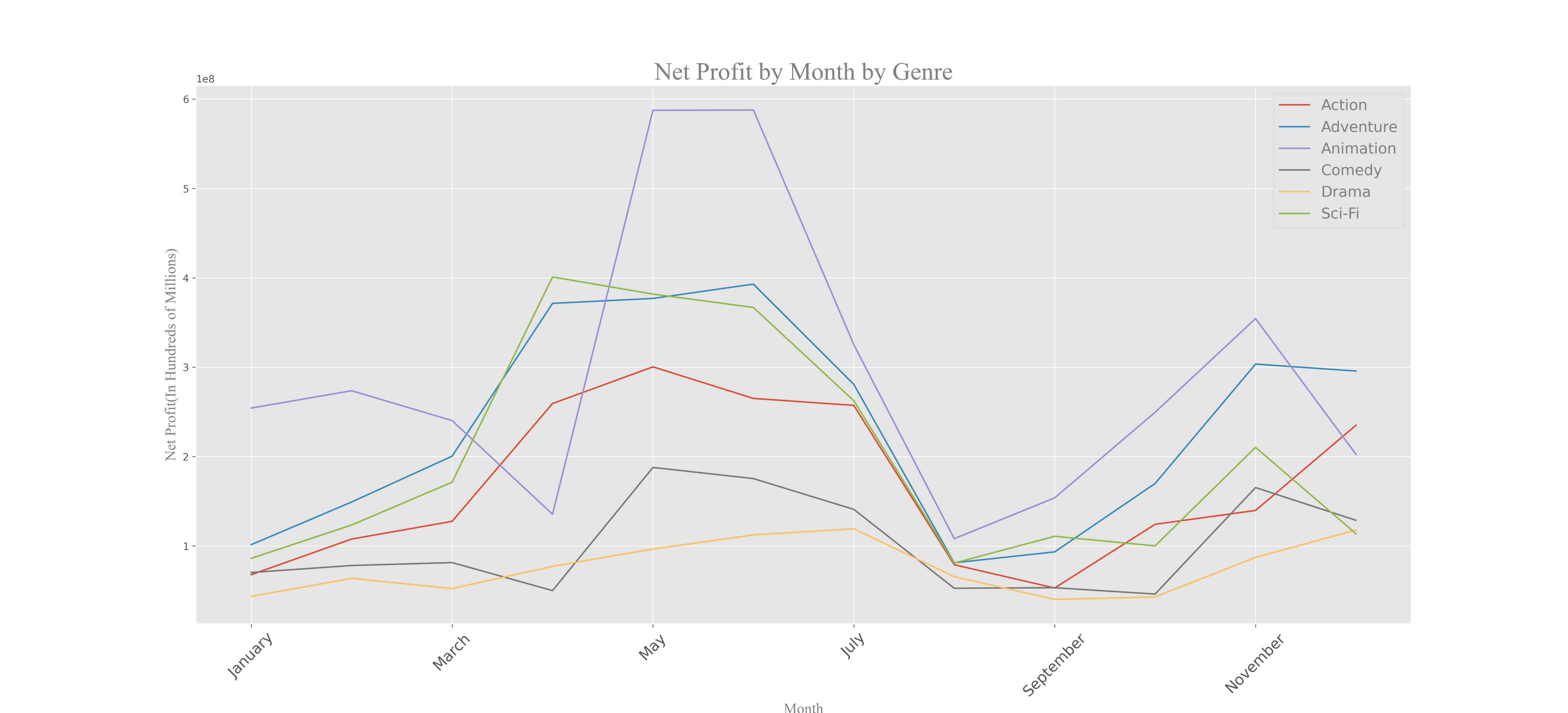
**Question 3: What is the best time of the year to release a movie?**

We start by converting the dates from the imdb\_budgets\_df dataframe to a datetime object. We then do a count by month to see the number of movies released in each month.

When grouping by month, we can select the Net Profit and Profit Margin columns so that we can see which months have the most financial success.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/MarginByMonth.png)

Finally we plot the net profit by month for a small selection of genres. We can see that there is a general trend amongst these genres for the profit in each month.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/ProfitbyMonthbyGenre.png)

**Question 3 Conclusion:** We recommend that Our Company release the bulk of their movies, especially Animation, during the summer months (i.e. May-July). Adventure, Drama and Comedy movies would see similar success if released in November, but the recommendation remains to focus on summer.

**Question 4: Which actors and directors tend to add the most value?**

In this section we are going to take a look at the average net profit across all movies. From there we want to determine which actors and directors consistently appear in movies where the net profit substantially exceeds the average. We will represent this in a field called Value Above Replacement(VAR). To further simplify this concept; if across all movies the average net profit is 100 dollars and the average net profit of movies from 'Actor: X' is 200 dollars he/she would have a VAR of 2. This number represents X times over the average. To eliminate outliers we will look at actors who appear in 10 or more movies and directors who work in 5 or more.

We'll use the actors\_df dataframe, adjust for inflation, and calculate profit as we did in Question 1. Then filter by actors who have starred in 10 or more movies.

actor\_counts = actors\_df['value'].value\_counts()

actor\_list = actor\_counts[actor\_counts >= 10].index.tolist()

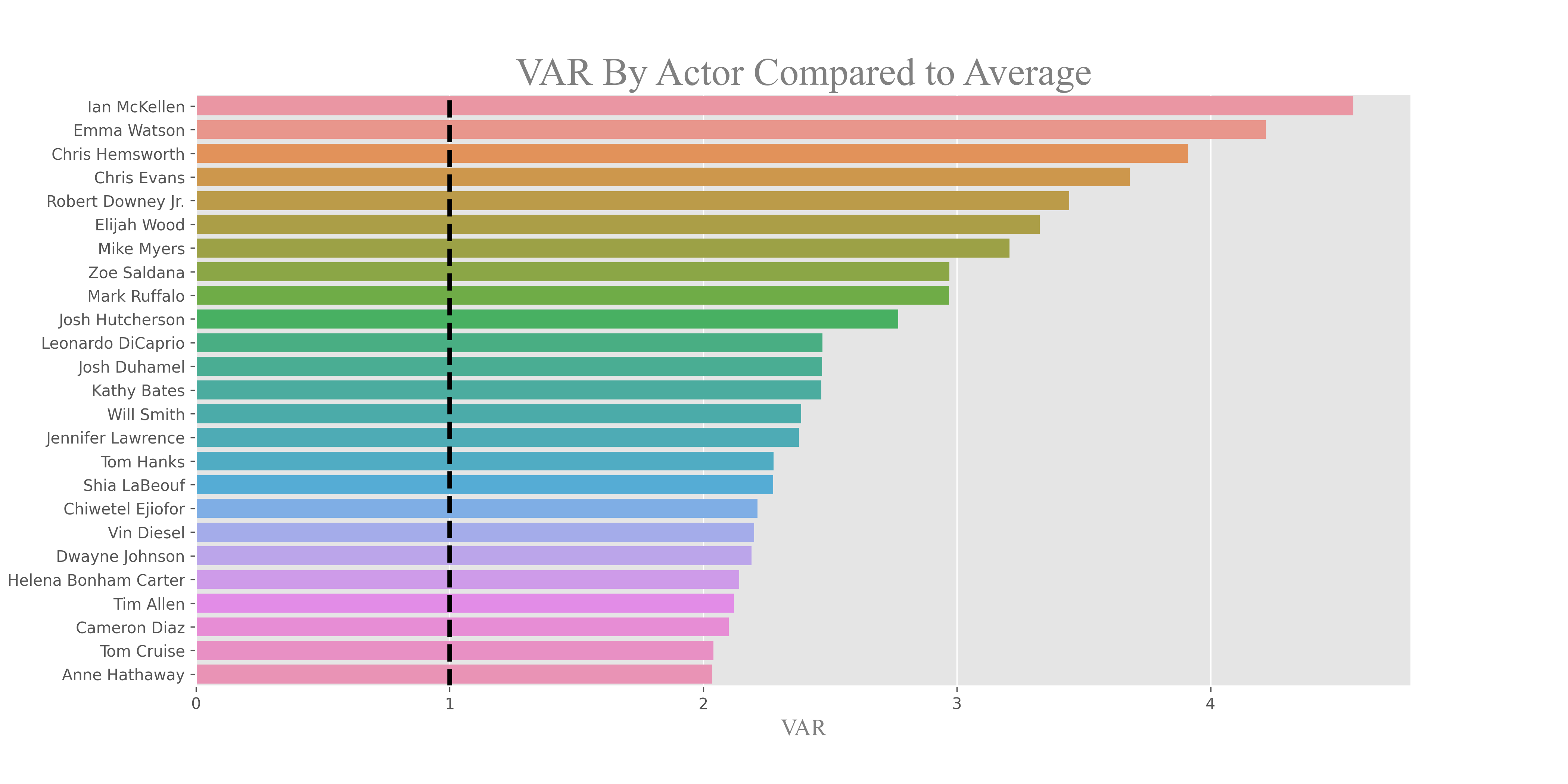
actors\_df = actors\_df[actors\_df['value'].isin(actor\_list)]

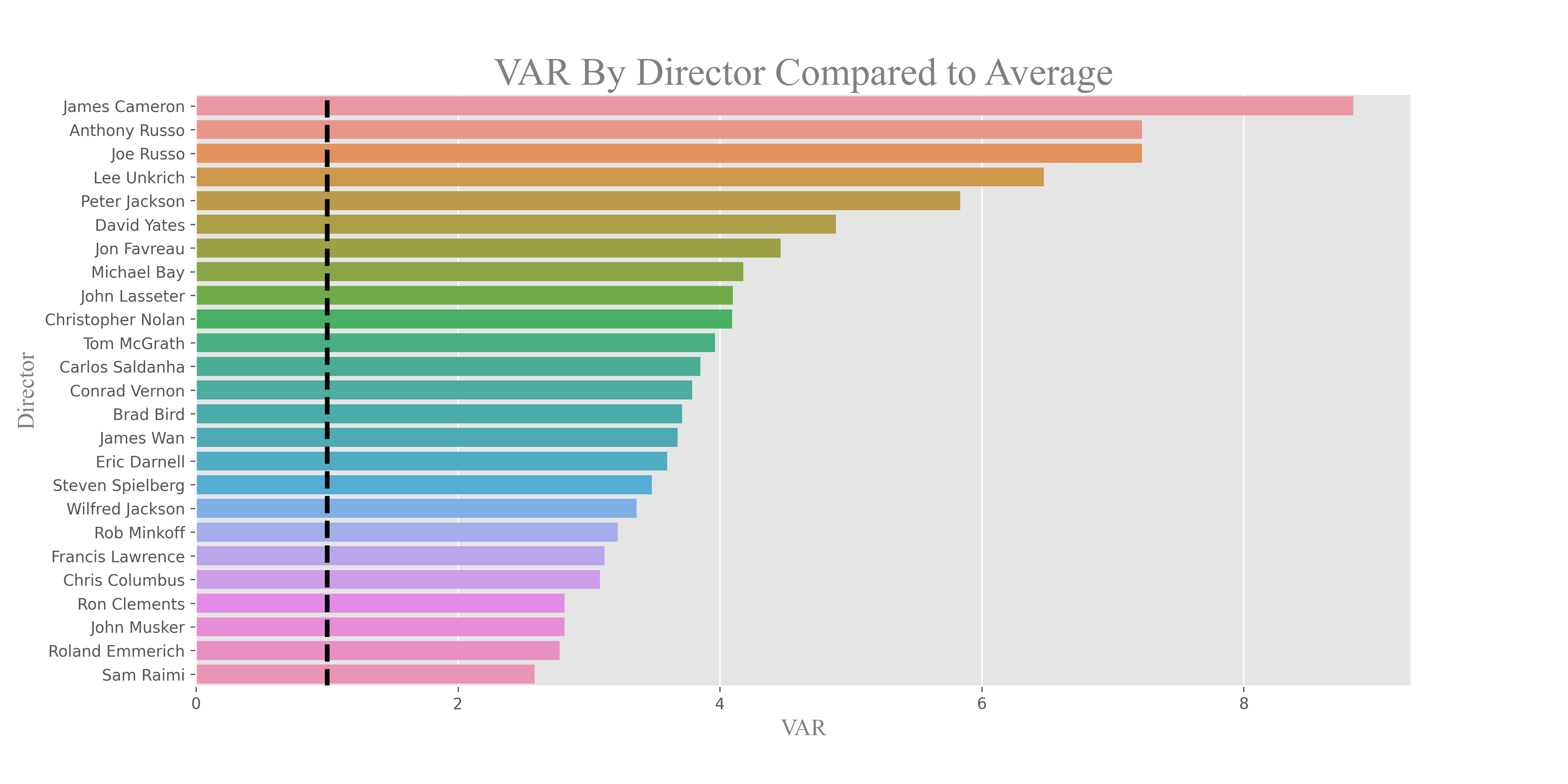
To calculate VAR:

actor\_total = actors\_df.groupby(['value'], as\_index=False)['Net Profit'].mean().sort\_values(by='Net Profit', ascending=False)

actor\_total['VAR'] = (actor\_total['Net Profit']/actor\_total['Net Profit'].mean())

We follow the same process with directors, except we filter by using 5 or more movies instead of 10.

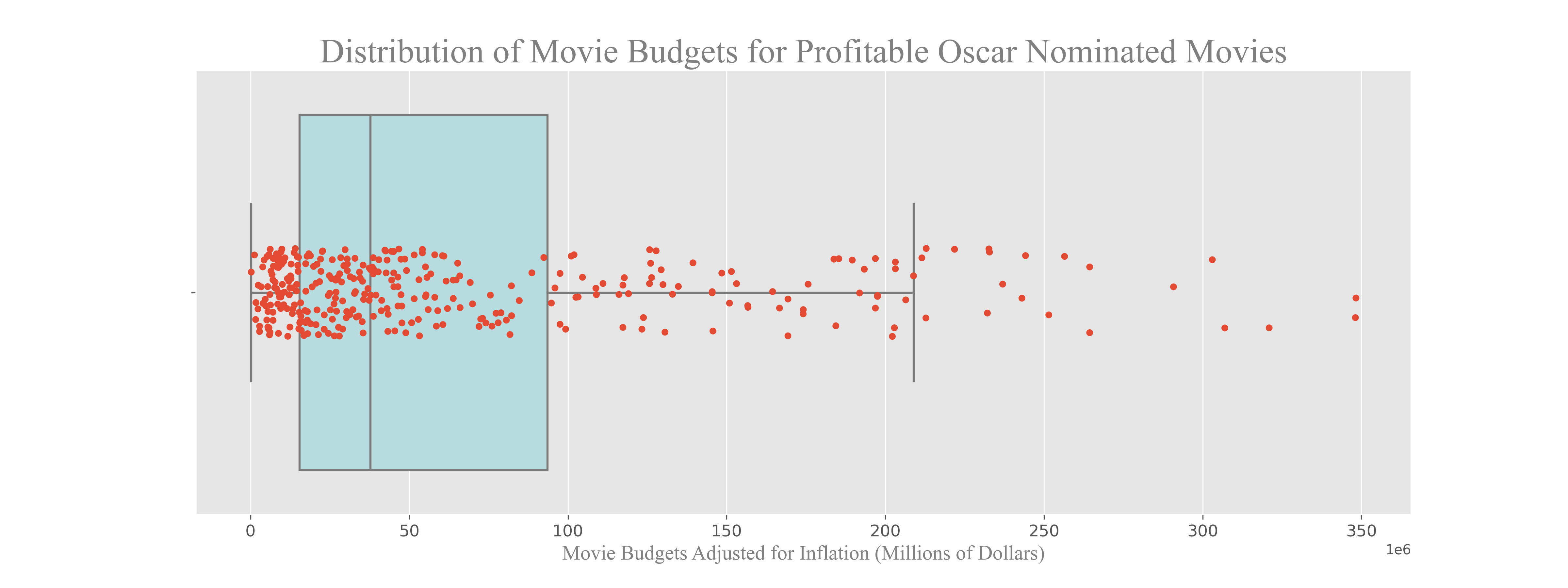
[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/VARActor.png)

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/VARDirector.png)

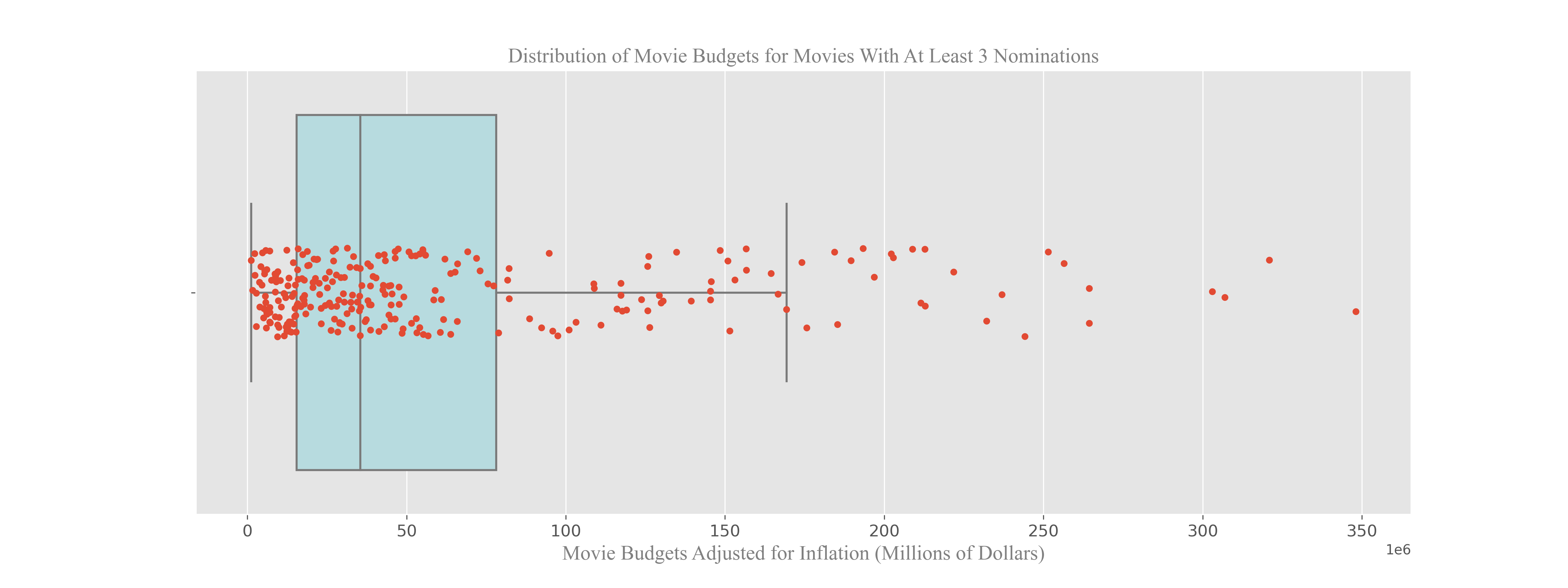
**Question 4 Conclusion:** We recommend that Our Company focus their cast and crew search to individuals who consistently score at least 1.0 on the VAR score. We can, with a high level of confidence, conclude that these individuals will elevate the overall production.

**Question 5: How much money should you spend on a movie to win an Oscar?**

In this analysis we will join the imdb\_budgets\_df and awards\_df dataframes so that we explore correlations between budgets and Oscar wins. We first look at a distribution of the budget for movies that have been Oscar-nominated. We also look at the win rates for the nominated movies so that we can establish the minimum number of nominations required to secure at least one win. For this analysis the minimum number or nominations required was three.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/Oscar_Nominated.png)

We use three nominations as the cutoff to filter our data and then look at the distribution of budgets again. We use the median as our measure of central tendency due to a large standard deviation in the data.

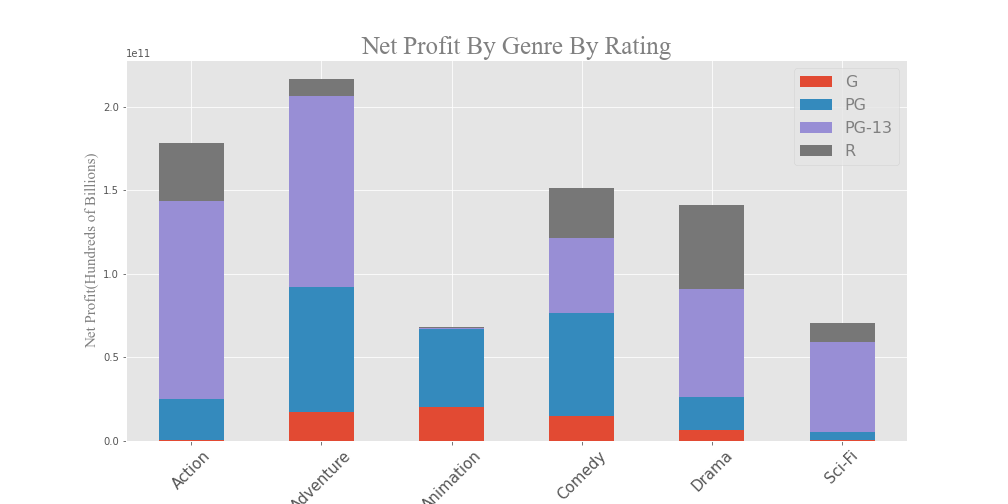
[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/3_Nominations.png)

**Question 5 Conclusion:** Our Company should spend at least $35,465,000 in order to make an Oscar-winning movie.

**Question 6: What impact, if any, does runtime and movie rating have on Net Profit, Profit Margin and IMDb rating?**

To answer this question we will only focus on the 4 ratings: G, PG, PG-13, and R. We then count the ratings to see how many movies fall within each category. From there we can examine the net profit and profit margin of genre to see which has the most financial success.

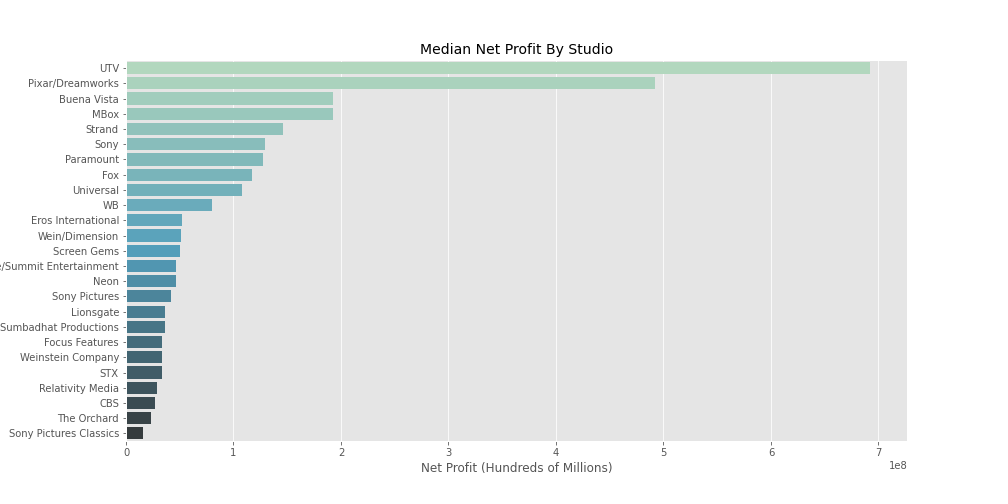
It's also important to see the net profits of each rating by genre. We first do a groupby on rating and genre and then create a pivot table so that we can see the net profits of each rating in each genre. This will guide us as to what ratings should be targeted based on the genre of the movie being made.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/ProfitbyGenrebyRating.png)

**Question 6 Conclusion:** We recommend that Microsoft take into consideration the rating of the movie based on the genre and target audience. If making animation movies, it is wise to stick to a G or PG rating, otherwise PG-13 is the sweetspot. In terms of runtime, there is little correlation in terms of overall profitability.

**Question 7: Sticking to our analysis of Net Profit and Profit Margin, what should Microsoft determine to be the baseline for sustainable success?**

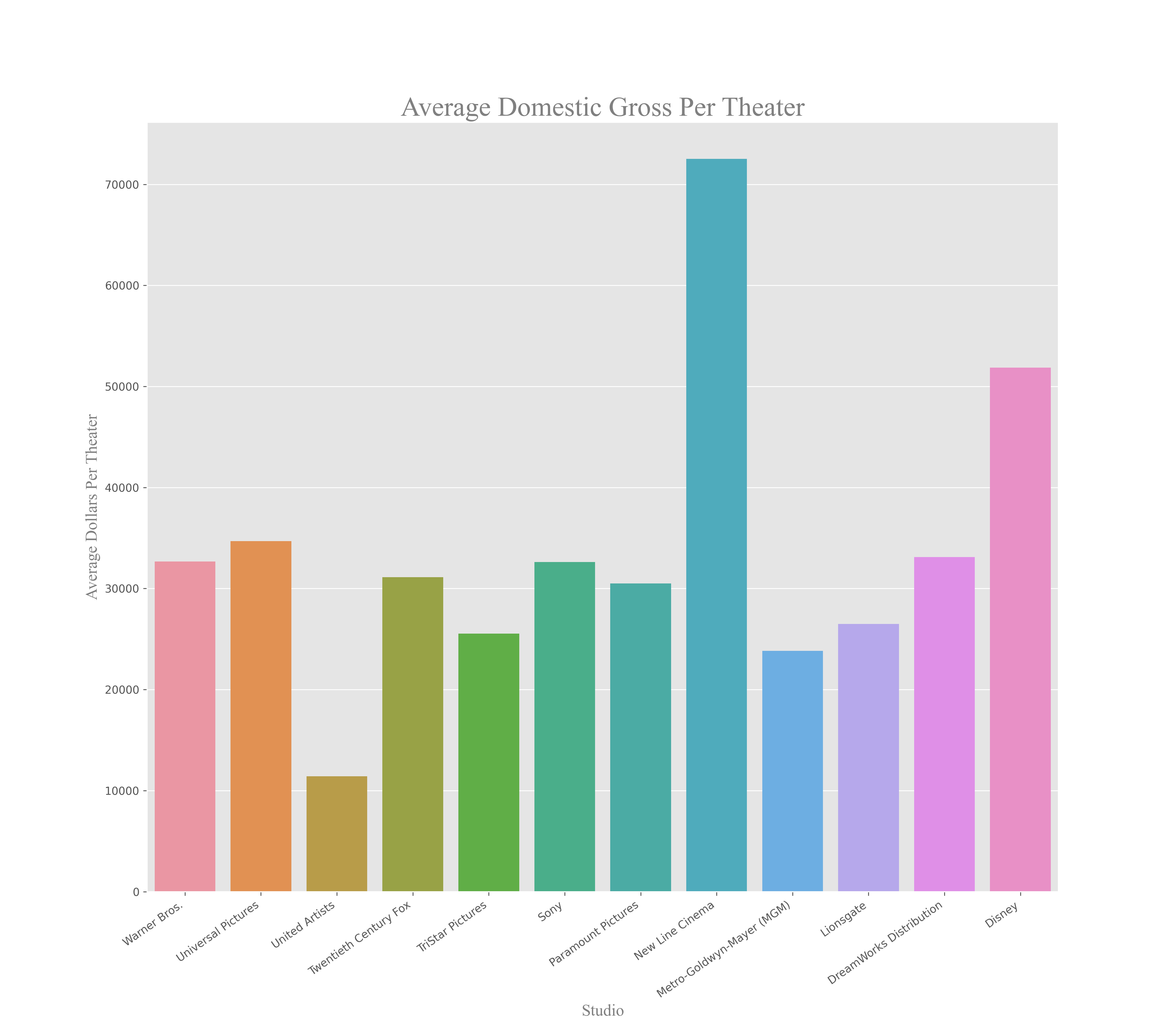
This analysis will require the use of the studiobudgets\_df dataframe. We first drop non-pertinent rows so that we can just focus on the studio performance. As we've done with other analyses, we use groupby and select the median as the primary measure of central tendency. We only select the top 25 studios as we are concerned with both being financial successful as well as being recognized as one of the major studios in the industry.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/NetProfitStudio.png)

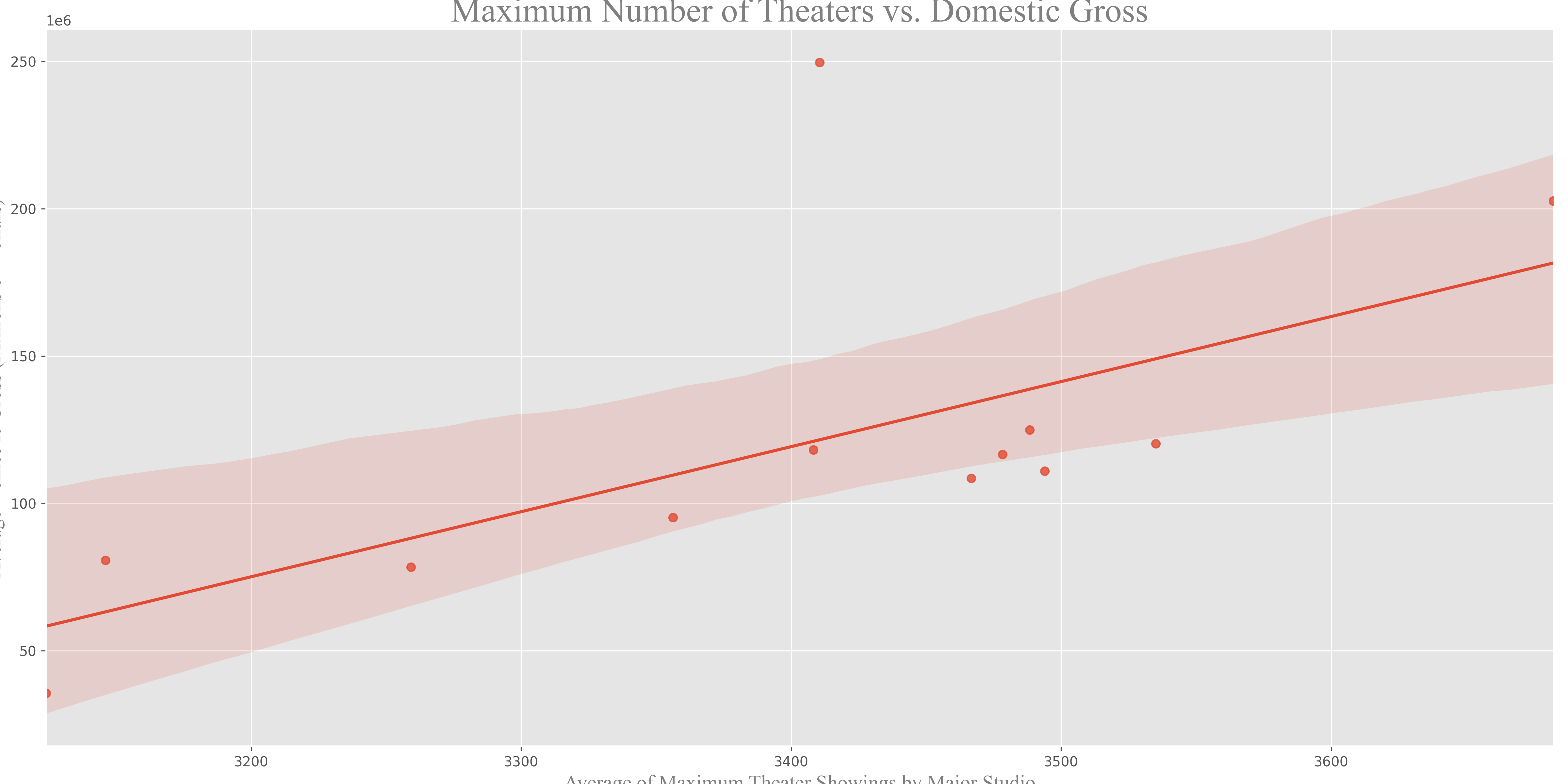
**Question 7 Conclusion:** Our Company should aim for a profit margin of 66% and a net profit of slightly over 50 million per movie to compete with the top existing studios.

**Question 8: Based on the success of current competitors, which should we look to for best practices?**

We take the theaters\_df dataframe and a column called dollars\_per\_theater to see what the average domestic gross per theater is for a movie. We then groupby by studio so that we can compare each studio.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/DomesticPerTheater.png)

We also investigate the correlation between the average maximum number of theaters and average domestic gross per studio.

[](https://github.com/PriyankaJhaTheAnalyst/Python_MoviesIndustry_EDA/blob/main/visuals/TheatersVGross.png)

We perform a join between the theaters\_df and awards\_df dataframes to further explore this question.

theaters\_df.set\_index(['title', 'year'], inplace=True)

theaters\_and\_awards = theaters\_df.join(awards\_df, how='inner', on=['title', 'year'])

The joining only leaves us with 66 movies upon which to draw conclusions. That is not ideal, but two studios - Disney and Warner Bros. - stand out among the rest as they have 22 and 15 of the movies in this dataframe, respectively. This is more than double any other studio. We then compare their average domestic gross per theater and win rate to see which of the two has performed better.

**Question 8 Conclusion:** Our Company should look towards Disney based upon domestic gross per theater and win rate. This means that Our Company should on average target for a movie to be in a maxium of 3818 theaters at its peak.