**What variables affect the later success or failure of a movie?**

Jonathan Hafez
MA213 C2

# Design of Study

## ABSTRACT

The topic I chose is based on ratings and variables of more than 10,000 movies, which are divided into many aspects. The dataset I used for this study is from Google Kaggle, which got it from The Movie Database. Throughout this study, I found very interesting relationships between different groups of variables which helped me prove hypotheses associated with this data: these are stated below. Additionally, I found other interesting relationships between different variables of related to movies production

## INTRODUCTION

In order to find a topic of study for this semester's research lab, I tried and found a dataset that has been influenced, in one way or another, by almost every person with access to a tv; that way, our results can represent all of these people. Therefore, I chose a dataset composed of many different movies from all over the world, with many different assessments for each movie. My connection to the movie industry is that it was a major part of my childhood, especially since I remember auditioning for movie parts. Nowadays, the movie & video production industry is worth approximately 25.8 billion dollar[1] as of 2022. In this industry, hundreds of directors and major companies are trying to determine the best actors and the best months to release their movies. Therefore, throughout this study, certain stereotypes will be tested, such as if, for example, the popularity of a movie does not influence the revenue. Additionally, other main factors will be considered; these include taking into account factors such as the budget, profit, and the popularity of the movie and deciding whether these factors are valid and if these values were part of the reason this movie was internationally successful or not.

**Our goal is to testify factors that influence the international success of a movie and include two hypotheses:**

- ***Null Hypothesis 1****: A movie's budget is of no relationship with the revenue produced.*
- ***Alternative Hypothesis 1:*** *Movies with a high budget tend to produce much more revenue than those with less budget.*
- ***Null Hypothesis 2****: A movie's revenue depends mostly on how popular the movie is. Very popular movies tend to have much higher revenue than those less popular.*
- ***Alternative Hypothesis 2:*** *The popularity of a movie does not influence the revenue. The revenue depends mainly on other variables.*

## EXPERIMENTAL METHODS

For this study, my sample is approximately 10,000 different movies that "The Movie DataBase" collected. Each row of this dataset is different from the others, containing the name and values of each respective value for that movie's column. While the columns, on the other hand, have different assessments for each movie. This assessment includes the revenue, the date it was released, and each movie's budget, among many others. All of these values have been

[1] Ibis World, main page.

slowly and accurately collected since 2008. Although we are looking at approximately 10,000 samples, not all of these will be considered due to missing values. An example of this is the movie “Pawn Sacrifice.” For this sample of the dataset, we are missing the budget this movie had, as well as the revenue. Like this sample, there are others with missing information. Thus because these samples are incomplete, these movies cannot be used to work as a sample to prove some of our assumptions.

This research will display different valuable information for almost each of these Movies. In order to get an accurate picture, our study will be broken down into various assessments for each particular movie. This dataset contains no demographic, as no specific group is being analyzed.

| ***Sample*** | ***Assessment of Different Aspects of a Movie*** |
|---|---|
| 10,900 different movies | • Budget<br>• Popularity<br>• Director<br>• Production Company<br>• Cast<br>• Revenue<br>• + More |

Table 1 – Assessment of the data

**METHODS**

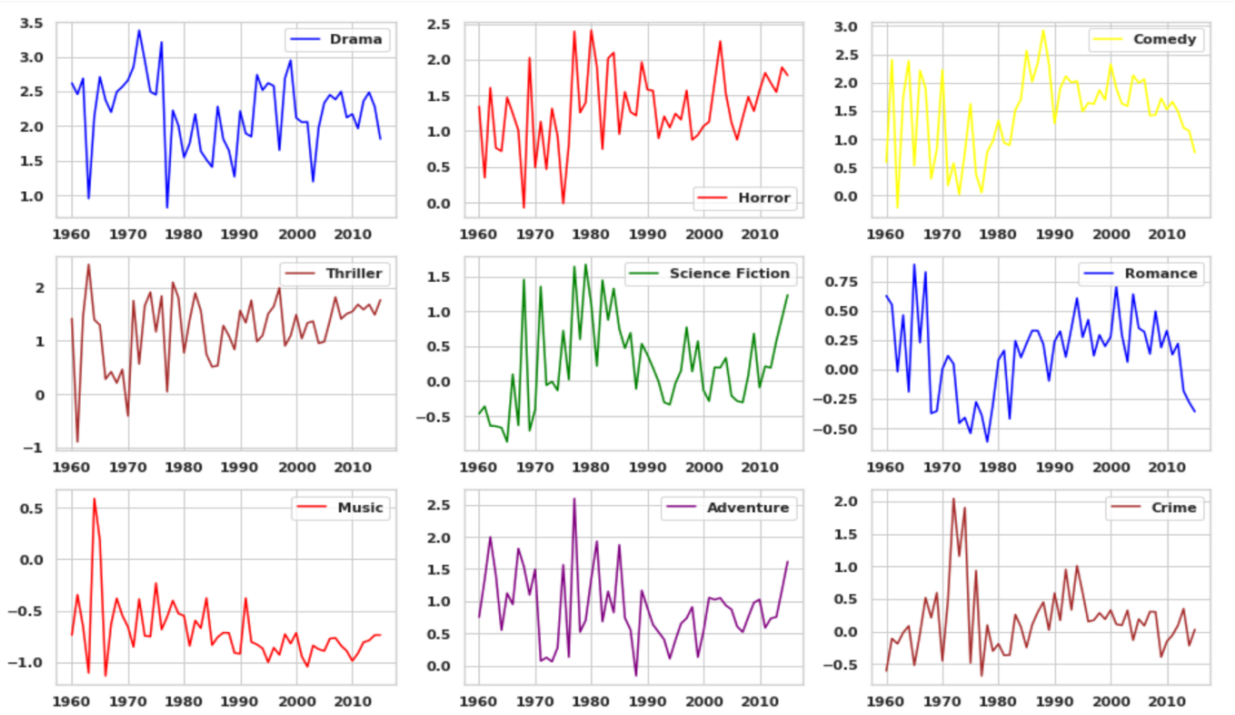
Data Collection:

Collecting data from so many movies is something that would take a long time. Therefore, I decided to use the public dataset that TMDB released to the public. It comes from the TMDB, which is short for The Movie Database. This shared community has been collecting data from movies back to 2008. Another company that also does the same is IMDb. However, the dataset that TMDB provides to the public is much more complete and has many more variables than that from IMDb. Even though the TMBD dataset is much cleaner, and fewer values are missing, the data had to be cleaned. To clean and understand the data better, we first had to analyze and decide what to do with specific missing values. After that, I created a new column called ‘Profit,’ which represents the revenue minus the budget. And lastly, I had to take the log for revenue, profit, and budget columns to have similar and comparable values. Creating a new column with the log of each of these respective columns allowed me to find a more accurate relationship between these and other variables, shortening the distance between big and small numbers. Almost all of the cleaning of the data, as well as many of the graphs, were done using python.

**Data visualization**

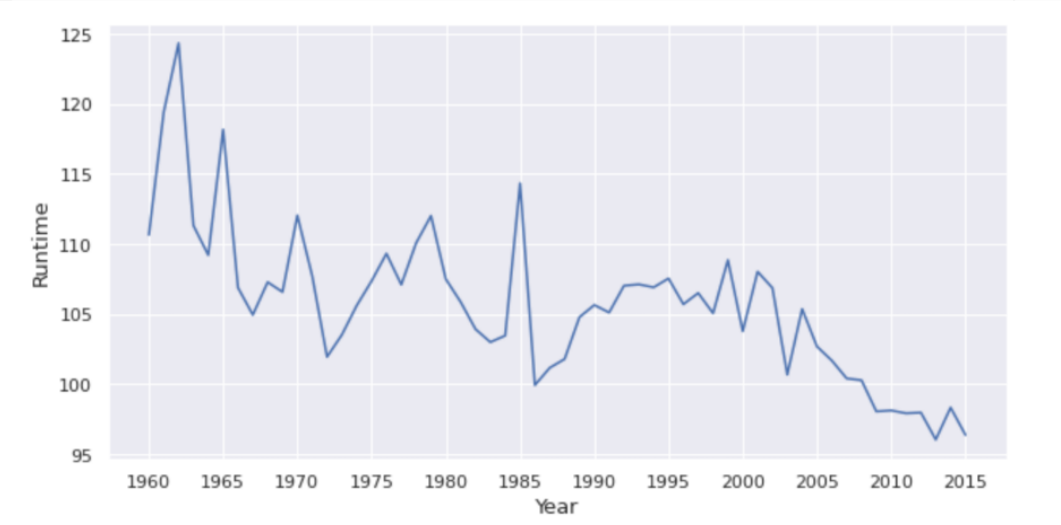
Throughout the study, data visualization was used to represent the relationships between different variables and to test the hypotheses. Through the use of various graphical techniques, such as bar charts, scatter plots, and line graphs, we were able to gain insights into the data and draw conclusions about the relationships between the variables. Additionally, data visualization allowed us to clearly and effectively communicate our findings to others, making it a significant part of this study.

*Movie Genre Popularity throughout the Years.*



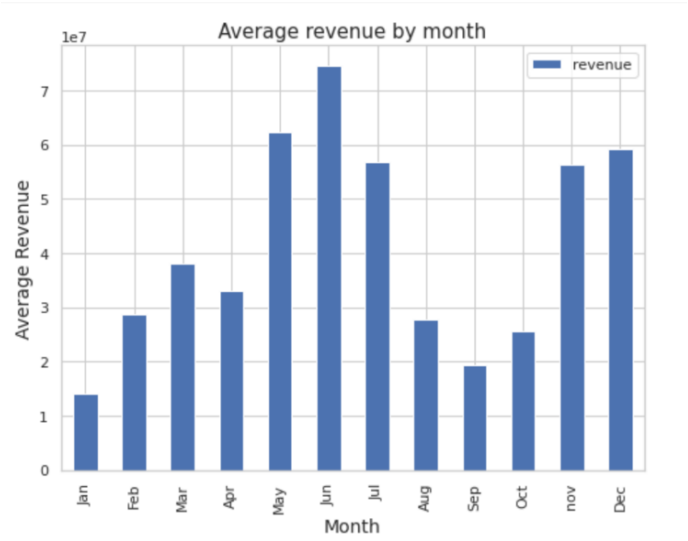

This plot was done by taking each genre's standard deviation and mean in each respective year. When the Z-Score are plotted, they allow us to see the relationship and popularity of each genre in each respective year. We can see that some genres are more popular than others, with specific genres consistently having higher Z-scores than others. This may be due to various factors, such as the changing tastes and preferences of audiences, the success of individual movies within a genre, and the quality of the movies within a genre. Overall, this plot provides valuable insights into the changing popularity of different movie genres over time.

## *Average Runtime of Movies*



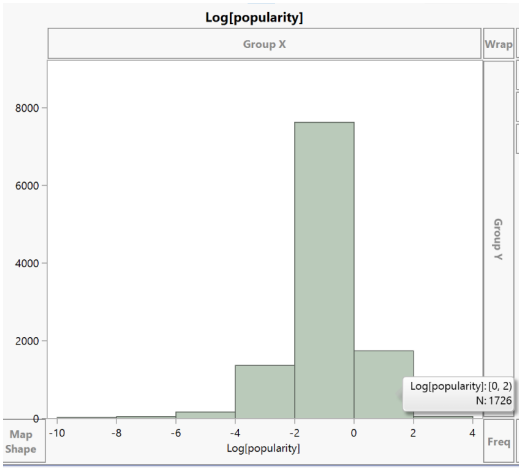

In this graph, we can see that the plotted line has a generally downward trend, indicating that the average runtime of movies has decreased over time. This may be due to various factors, such as the increased use of technology in movie production, which allows for more efficient editing and pacing of the movie. Additionally, the increasing demand for content and the abundance of streaming platforms may also be contributing to the trend of shorter movies. Overall, this graph provides useful insights into the changing trend of movie runtime over time..

## *Average Revenue by Month (1960-2022)*



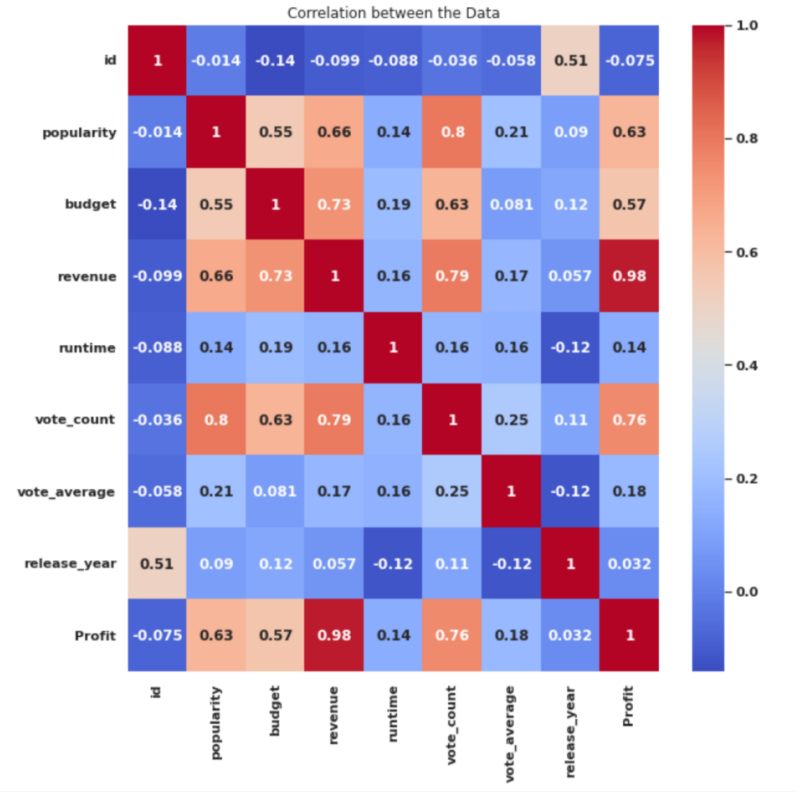

In this bar chart, we can see that the bars have a generally upward trend, with the highest bar occurring in the summer months of June. This indicates that movies released in the summer generate more revenue than movies released at other times of the year. One possible explanation for this result is that many people take vacations during the summer, which gives them more time to attend the movie theater. Additionally, the summer months are also when many blockbuster movies are released, which may also contribute to the higher revenue generated by movies released in the summer. Overall, this bar chart provides valuable insights into the relationship between the month a movie is released and the revenue it generates.

## *Log(Popularity)*




This bar chart shows the distribution of popularity scores for the movies in our data set. To better visualize the distribution of the data, the log of each popularity score was taken, which transformed the data into a more compact and symmetrical form. This allowed us to see the overall pattern of the data without the influence of extreme values or outliers, which can often distort the shape of a distribution.

## *Correlation Matrix*




In this correlation matrix, we can see the relationship between different variables in our movie data set. For example, the “Profit” and “Revenue” variables are highly correlated, with a coefficient of 0.98. This indicates that there is a strong positive relationship between these two variables and that movies with higher revenues tend to have higher profits.

Rapid Increase in the Production of Movies.

In figure 1, we can see that as the years have passed, the number of movies produced is growing yearly. There are two possible explanations for the increase in movie productions during the last couple of years. The first explanation is that as time went on, The Movie DataBase started to grow, and because they have a more extensive community, it can capture all of the different assessments for many recent movies. A second and more valid explanation is the rapid increase in technology and the new streaming platforms allowing us to stream movies anywhere. These changes have made the movie industry much more valuable than it previously was, and this popularity has increased the number of movies produced.

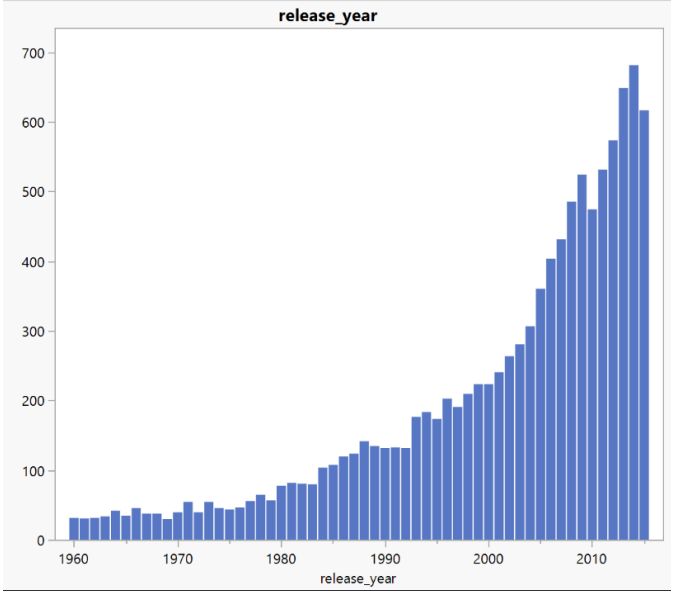



Figure #1

In this figure, we can see that the bars have a generally upward trend, indicating that the number of movies produced has increased over time. The plotted bars show that the rate of growth has increased over time, with the bars becoming taller in more recent years. This suggests that the movie industry has been rapidly expanding and that the number of movies produced is likely to continue increasing in the future.

To make the movie industry profitable enough, many movies must be produced. This can be seen from the relationship between the two graphs. In exhibit A, one can see the name of the production companies with the most movies produced. While in exhibit B the profit of the production companies is shown. Most of the production companies' names in both graphs are the same. Thus, the biggest production companies tend to be more profitable than others. And the reason they are more profitable may not be because their movies are better and have a higher level of popularity but because they produce more movies than others.

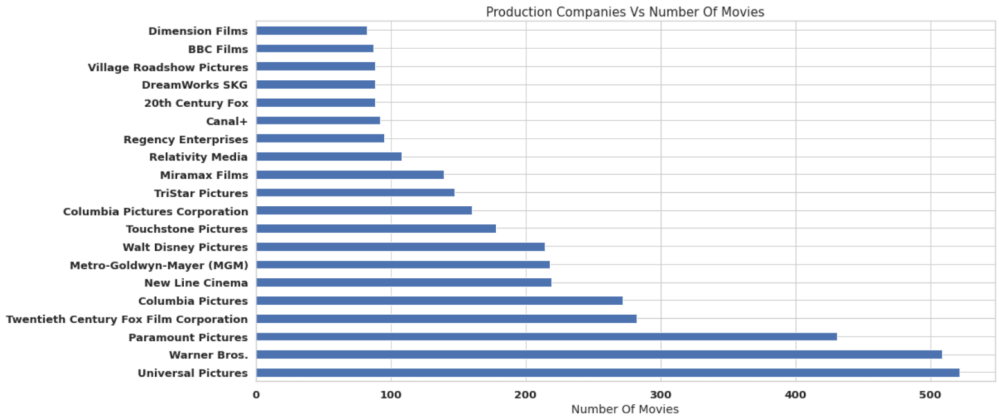



Exhibit A

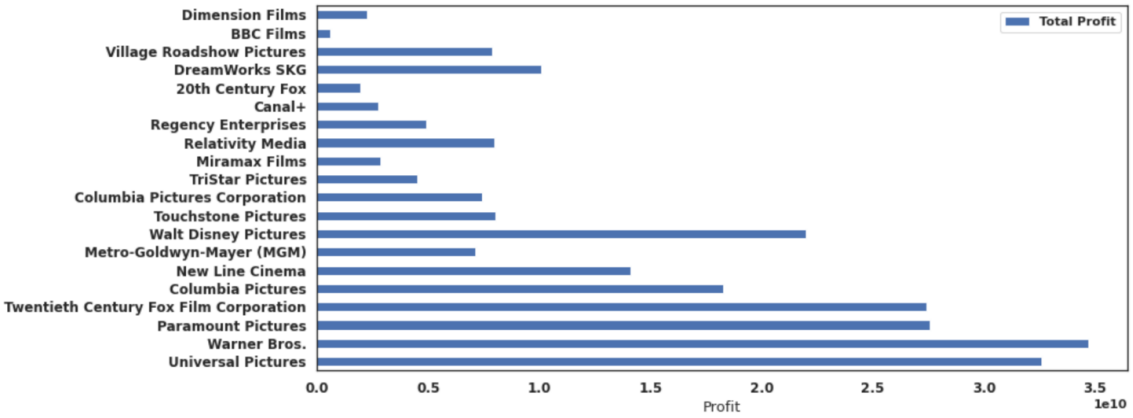
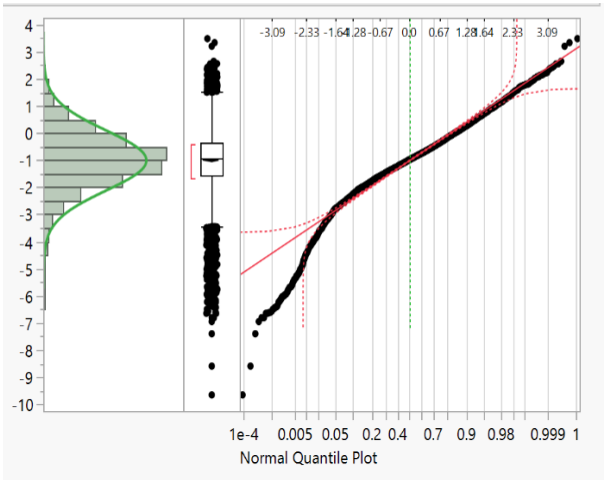


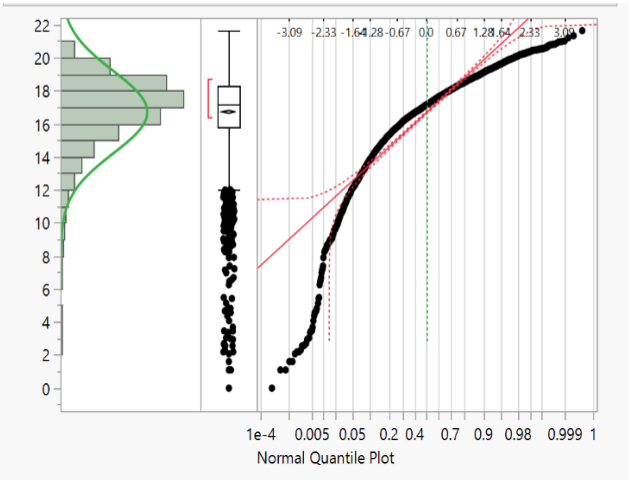

Exhibit B

---

## *Assessment of Normality*




***log(Popularity)***




***log(Revenue)***

The assessment of normality is a statistical method used to determine how well a given data set conforms to a normal distribution. This is important because many statistical tests assume that the data follows a normal distribution, and non-normal data can lead to invalid or misleading results. In this study, we assessed the normality of two variables: log(Popularity) and log(Revenue). We used quantile plots and goodness-of-fit tests to visualize the data distribution and evaluate whether it is consistent with a normal distribution. When plotted, I noticed that the log(revenue) is not within the confidence bands. (The confidence interval is at a 95% significance level for both of these variables, and the sample size is very large with n= 10866). Because they are not in the confidence bands, we cannot conclude that our data is normal. However, on the log(popularity), we can say that because all of the points are either within or touching the confidence bands, we cannot conclude that our data is non-normal.

---

## Hypothesis #1

- ***Null Hypothesis 1****: A movie's budget is of no relationship with the revenue produced.*
- ***Alternative Hypothesis 1:*** *Movies with a high budget tend to produce much more revenue than those with less budget.*

The first hypothesis of this study can be proven by the results of our data. As shown in figure 3, it can be seen that many movies with high budgets had more revenue than those with less funding. Because R^2 is .54, a movie's budget could be one of the ingredients for generating high revenues in the movie industry. Looking at our correlation matrix, we can see a very high correlation between the budget and the revenue, .73. Revenue is the variable with the most significant correlation with budget. ( the "revenue" is not the movie's total profit; the budget must be decreased from the revenue to get the profit).

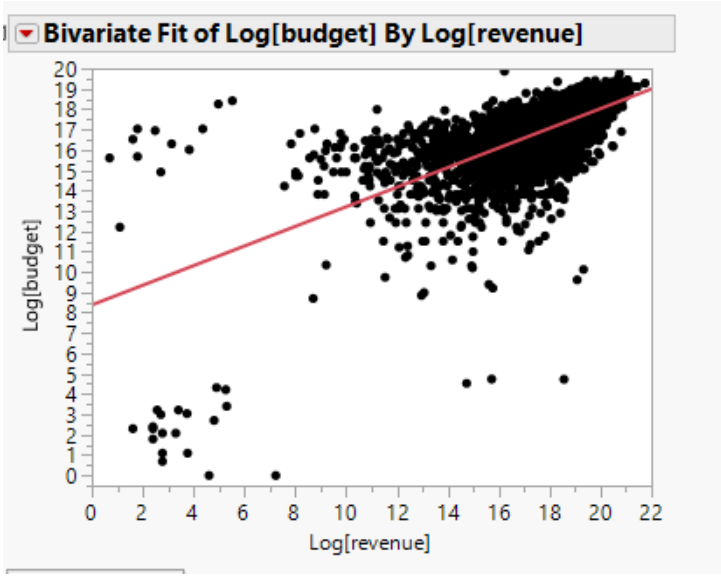
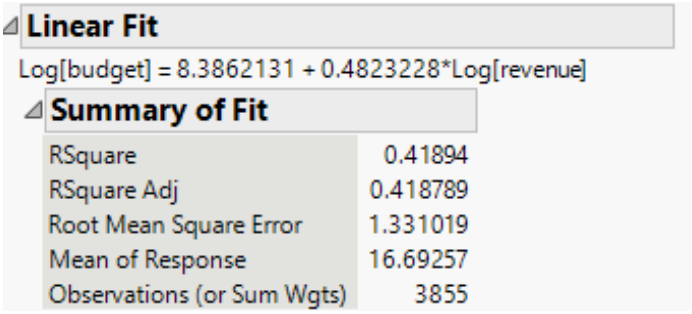



Figure 3.

This figure shows that the fitted line has a positive slope, indicating that movies with higher budgets tend to have higher revenues. This is consistent with our alternative hypothesis. Additionally, the scatterplot shows that there is a lot of variation in the data, with some movies having very high revenues even with relatively low budgets and vice versa. This could suggest that budget is not the only factor influencing a movie's revenue and that other factors may also play a role.




**T - Test** -Here, The t-test will be performed to see if the budget has any significant relationship with the revenue.

**Hypothesis Testing #1**

$$t = \beta/SE\beta = 8.3862/1.331 = 6.301$$

The p-value is < .00001.

The result is significant at p < .05.

In conclusion, after analyzing the data, the null hypothesis is rejected, as sufficient evidence indicates that high budgets are essential for high revenue numbers. In the t-test, the calculated t-value is 6.301, which is much larger than the critical t-value at the 0.05 significance level. This indicates that there is a very low probability that the observed relationship between the budget and the revenue occurred by chance. Since the p-value is also less than 0.00001, it is clear that the test result is statistically significant, and the null hypothesis can be rejected in favor of the alternative hypothesis.

---

## ***Hypothesis #2***

- ***Null Hypothesis 2****: A movie's revenue depends mostly on how popular the movie is. Very popular movies tend to have much higher revenue than those less popular.*
- ***Alternative Hypothesis 2:*** *The popularity of a movie does not influence the revenue. The revenue depends mainly on other variables.*

As we can see in the results, how popular a movie is, has a clear correlation with its revenue. The correlation between these two variables is .66; therefore, we can say they are linearly dependent. Yet, budget, on the other hand, does not have that much of a correlation with popularity. This can be seen in many cases, one of those examples is the movie John Carter. Before its release in 2015, this movie had very high expectations due to its very high production cost, approximately 263 million dollars. When released, it did not live up to its expectations and did not become popular. The production company lost between 133 and 240 million dollars. Although the budget is not a direct trait for success, popularity is a trait of revenue (Figure #5).

---

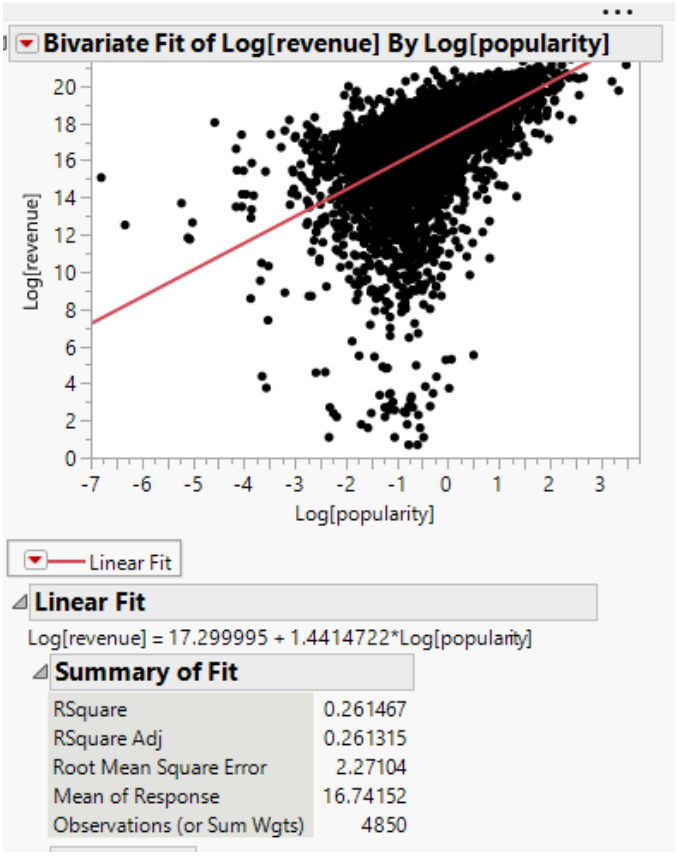



Figure #5

From the plot, it can be seen that as a movie's popularity increases, its revenue also increases. This suggests that there is a positive relationship between these two variables and that movies that are more popular generate more revenue than those that are

less popular. This relationship is further supported by the positive correlation coefficient of 0.66, which indicates a positive relationship between popularity and revenue.

**Hypothesis Testing #2**

$$t = b/seb = 17.299/2.27104 = 7.617$$

The p-value is < .00001.

The result is significant at p < .05.

To conclude, the t-test indicates a strong relationship between a movie's popularity and revenue. The calculated t-value of 7.617 is much larger than the critical t-value at the 0.05 significance level, which means that it is very unlikely that the observed relationship happened by chance. The p-value is also less than 0.00001, supporting the conclusion that the relationship is statistically significant. As a result, the null hypothesis that the popularity of a movie does not influence its revenue can be rejected in favor of the alternative hypothesis that the popularity of a movie does influence its revenue.

---

## Conclusion

In conclusion, this study analyzed a dataset of over 10,000 movies to test two hypotheses about the factors that influence the international success of a movie. The first hypothesis was that a movie's budget has no relationship with its revenue, while the alternative hypothesis was that movies with a high budget tend to produce more revenue. The second hypothesis was that a movie's revenue depends mainly on its popularity, while the alternative hypothesis was that the popularity of a movie does not influence its revenue. Through the statistical analysis, the study found that a movie's budget and popularity were strongly correlated with its revenue and that the null hypotheses for both were rejected in favor of the alternative hypotheses. This indicates that movies with high budgets and high levels of popularity tend to produce more revenue than those with lower budgets and lower levels of popularity. Furthermore, the study found that the director, production company, and cast of a movie were also correlated with its revenue, though to a lesser extent than budget and popularity. These findings provide valuable insights for movie studios and production companies looking to maximize the revenue of their films and make informed decisions about the budgets, marketing efforts, and personnel for their movies. Additionally, the study highlights the importance of considering a range of factors when predicting the success of a movie and suggests that future research could further investigate the specific mechanisms through which budget, popularity, and other factors influence a movie's revenue. Overall, the study contributes to our understanding of the factors that drive the success of a movie in the international market and has practical implications for the movie industry.