Movie Industry Insights: Making Blockbusters with Data

李兆智 Li Zhaozhi

BUAN 3065 – Marquette University

November 7, 2020

Background

Movies have been a mainstream form of entertainment for decades. Production houses compete to produce box office hits. It is critical to understand what elements make movies successful. To do this, we need to investigate some questions by analyzing historical data:

1) Which genre is most likely to produce the highest gross?
2) Which rating is most likely to produce the highest gross?
3) What is the relationship between gross and IMDb scores?
4) What is the relationship between gross and the popularity of movie stars?

Genres and ratings are determinantal of the demographics of the target audience. I expect some genres and ratings to produce higher revenue than others. Many people consult scores and the information of featured stars to select movies. I expect movies with higher revenue to have higher scores and more popular featured stars.

Introduction

The following Data Definitions Table contains the variables in the dataset, the definition of each variable, and the sources of this data.

Table 1. Data Definitions Table

Variable Name	Variable Definition	Source
budget	Budget of the movies in United States Dollars.	IMDb
company	Name of the production company	IMDb
country/region	Place of origin of the movie	IMDb
director	Movie director's name	IMDb
genre	Genre of the movie	IMDb
gross	Revenues of the movie in United States Dollars	IMDb
name	Name of the movie	IMDb
rating	Rating of the movie (G, PG, R, etc.)	IMDb
released	Date of theoretical release	IMDb
runtime	Length of the movie	IMDb
score	Score of the movie	IMDb
star	Name of the featured movie star	IMDb
votes	Number of votes the star receives	IMDb
writer	Name of the storywriter	IMDb
year	Release year of the movie	IMDb

IMDb is an online movie database with active users from across the world and is one of the most frequently quoted movie data sources. There are 4576 observations from the period between 1986 and 2016 in this dataset. These observations cover 44 countries and regions with most of them coming from the United States.

Revenues are an appropriate measure of a movie's success because the goal of movie studios is to maximize profits. Scores and votes are the most direct ways for audiences to express their opinions on movies and movie stars. These measures will help answer those questions.

Descriptive statistics can help prepare the foundation for more in-depth analyses as we try to retrieve insights from this data. The following Quantitative Measures Table contains some descriptive statistics.

Table 2. Quantitative Measures Table

	budget	gross	runtime	score	votes
Mean	$36,145,601.60	$46,074,694.34	107.60	6.36	95702.54
Median	$23,000,000.00	$23,455,506.50	104.00	6.40	43940.00
Mode	$20,000,000.00	$20,100,000.00	100.00	6.70	5757.00
Standard Deviation	39969473.26	66293784.14	18.02	1.01	149387.83
Minimum	$6,000.00	$309.00	69.00	1.50	183.00
Maximum	$300,000,000.00	$936,662,225.00	280	9.3	1861666
Coefficient of Variation	111%	144%	17%	16%	156%
25th percentile	$10,000,000.00	$6,290,004.75	95.75	5.80	16107.50
75th percentile	$46,000,000.00	$57,814,795.50	117.00	7.10	109462.50

The coefficient of variation is a statistical measure of variability of quantitative variables. We can see this value is high for budget, gross, and votes – they are all well above 100%. High volatility is considered risky in finance. We can see the coefficient of variation of gross is even higher than that of budget. We can say the movie business is highly volatile.

Median is the middle number in an ordered list, which is a useful measure. From the median of budget, we can see many movies cost around 23,000,000 USD to produce. The median of gross is 23,455,506 USD, which is not much higher than that of budget. This comparison indicates that it is challenging to make money from the movie business.

We can also see the volatile nature of the movie business by looking at the minimums. The minimum is the smallest number. The lowest gross is 309 USD, which is not an error but an outlier. Many sources confirmed this three-digit box office sales number. My research shows that the movie is called *Trojan War* and costed around 150,000,000 USD to make. It is one of the worst market performers in history. While this is not representative of the industry, it is a powerful reminder of how bad a failure can go.

An overview of categorial data will help us understand this data even better. Frequency tables are helpful for this analysis. They should tell something about the movie industry at a high level.

Table 3. Frequency Table – Genre

	Genre	Frequency	Percentage Frequency
1	Action	1094	23.91%
2	Adventure	289	6.32%
3	Animation	229	5.00%
4	Biography	234	5.11%
5	Comedy	1292	28.23%
6	Crime	350	7.65%
7	Drama	770	16.83%
8	Family	6	0.13%
9	Fantasy	28	0.61%
10	Horror	227	4.96%
11	Musical	1	0.02%
12	Mystery	30	0.66%
13	Romance	5	0.11%
14	Sci-Fi	11	0.24%
15	Thriller	8	0.17%
16	Western	2	0.04%
	Total	4576	100%

Table 3 is the frequency table for genres. There are 16 genres in this dataset. We can see most movies during this period are considered comedies. This genre has the highest frequency, which is about 28% of all movies. Action is the next, and drama is in the third place.

These data are meaningful because we can see the popularity of each genre. Companies are most likely to produce movies with the most desirable genres. This should serve as a guideline for movie studios when they pick stories.

Table 4. Frequency Table - Rating

	Rating	Frequency	Percentage Frequency
1	G	100	2.19%
2	NC-17	9	0.20%
3	PG	659	14.40%
4	PG-13	1561	34.11%
5	R	2247	49.10%
	Total	4576	100.00%

Table 4 is the frequency table for rating, another measure we will use to answer one of the questions. We can see there are five rating levels in this dataset.

The Motion Picture Association created this movie rating system to rate a movie's suitability for certain audiences based on its content. This is important for producers to know because certain

ratings will pose age restrictions on audiences and impact box office sales and many aspects of the business.

From Table 4, we can see most movies are R-rated and only a few are NC-17-rated. R-rated movies almost counted for 50% of all movies. These two ratings are meaningful. R-rated movies are not suited for audiences under the age of 17. Audiences under this age require parental guidance to watch these movies. NC-17 strictly deprives audiences under 17 of watching the movies.

These data suggest that while most movies are filmed for adults, only a few strictly prohibit audiences under a certain age. PG-13 is a more desirable rating as it allows all family members to watch. This rating counts for 34% of all movies released during the period and is in the second place according to the frequency table. This information will be helpful as we use the data to retrieve insights.

The correlation coefficient plays a pivotal role in descriptive analytics. It will help us see the relationships between quantitative variables. Table 5 is a Correlation Coefficient Table.

Table 5. Correlation Coefficient Table

	Parameter1	Parameter2	correlation coefficient	t	p-value
1	budget	gross	0.68	62.37	0
2	budget	runtime	0.32	22.70	1.31E-107
3	budget	score	0.08	5.35	9.47E-08
4	budget	votes	0.45	33.99	3.33E-225
5	gross	runtime	0.26	18.09	3.11E-70
6	gross	score	0.24	16.40	1.76E-58
7	gross	votes	0.64	56.58	0
8	runtime	score	0.42	31.07	1.17E-191
9	runtime	votes	0.37	26.54	9.28E-144
10	score	votes	0.48	36.64	5.04E-257

Table 5 contains correlations for all quantitative variables in this dataset. A negative correlation indicates that the two variables move in opposite directions whereas a positive correlation indicates that the two variables move in the same direction. A correlation close to 0 indicates that the two variables are relatively independent. A correlation should fall between -1 and 1. The t and p values are useful for measuring the level of significance. They will help interpret the results.

We are interested in gross, which indicates the movie revenue. The table indicates that the budget and gross share the highest correlation, which is 0.68. This indicates that movies with heavier investments are likely to generate higher revenue.

Votes and gross also share a high correlation, which is 0.64. Votes indicate audiences' support for movie stars. This indicates that movies featuring popular stars are likely to generate higher revenue.

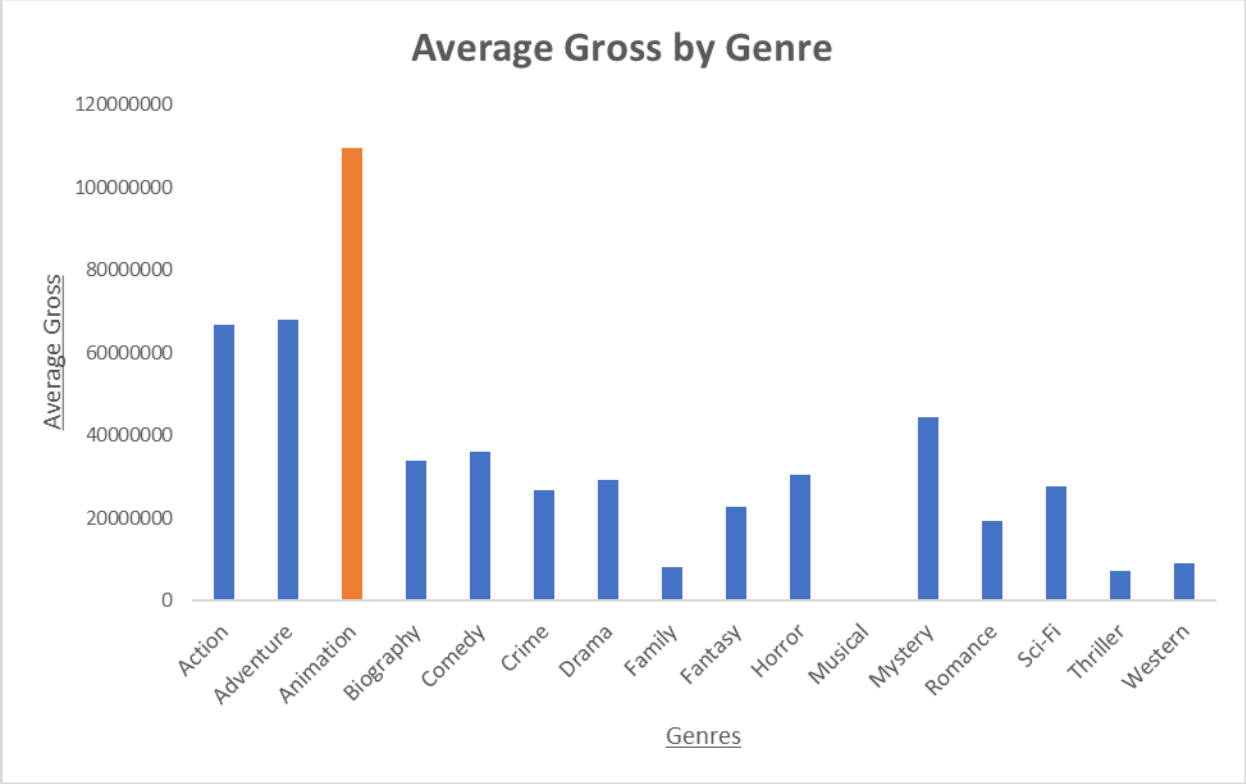
The corresponding p-values to both are 0. The p-values of 0 suggest that we reject the null hypotheses, which also suggest that there are correlations between these variables.

Insights

Now having explored this data using statistics and retrieved some insights, we are ready to find answers to those questions.

The first question concerns the relationship between genres and revenue. It should help answer this question by looking at average gross by genre. Graph 1 is a column chart created using genres and gross.

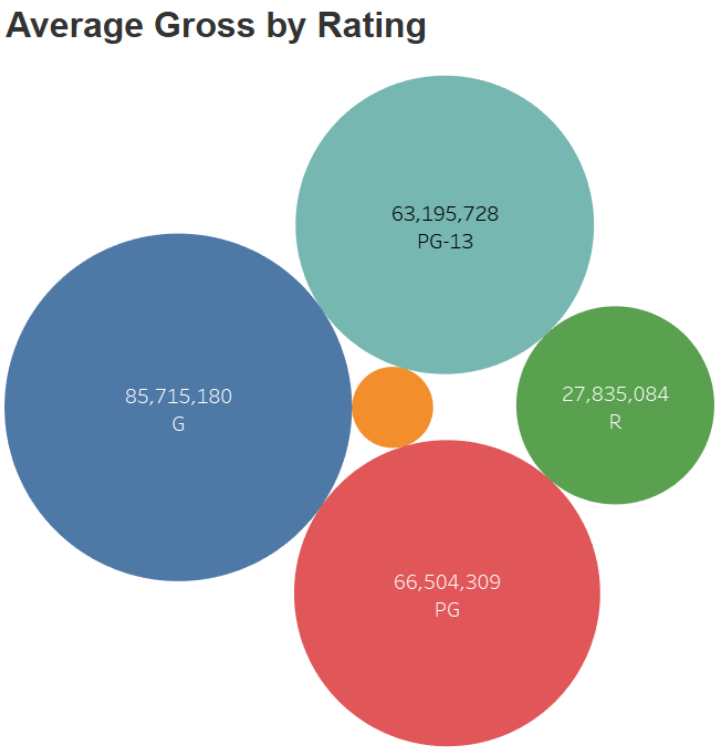
Graph 1 Average Gross by Genre in USD



The bar chart clearly indicates that the animation genre generated the highest gross. This is an interesting finding because Table 3 indicates that animation was not in the top three during the period. Movie studios should consider producing animated feature films to maximize profits.

Ratings will also impact movie revenue. The second question asks for the relationship between ratings and revenue. We can make a bubble chart this time to visualize the relationship between these two variables. Graph 2 shows the average gross by rating.

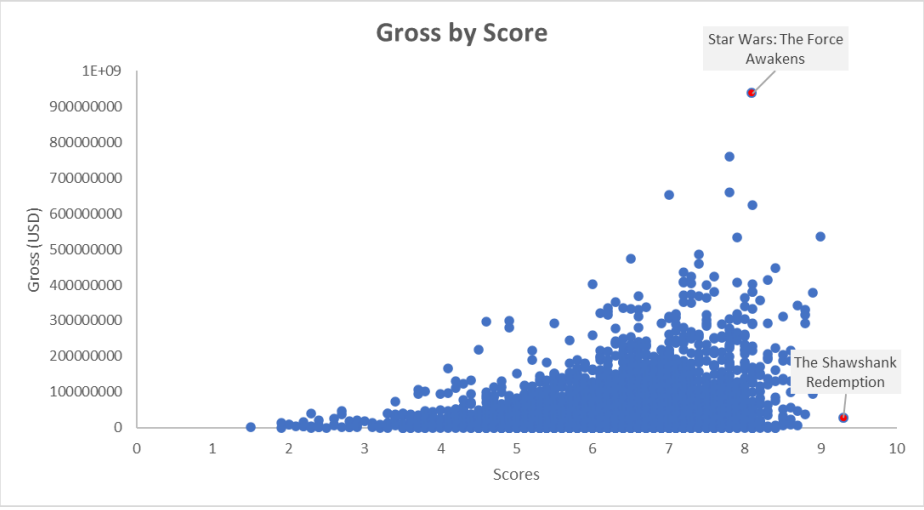
Graph 2. Average Gross by Rating in USD



The bubble chart indicates that G, PG, and PG-13 are the top three ratings that generated the highest average gross. Combining with our previous finding, movie studios should produce G-rated animated movies.

The relationship between gross and scores is also our interest. IMDb users give scores to each movie using the scale of 1 to 10. IMDb calculates the average score for each movie. We can visualize this relationship by making a plot chart using the two variables.

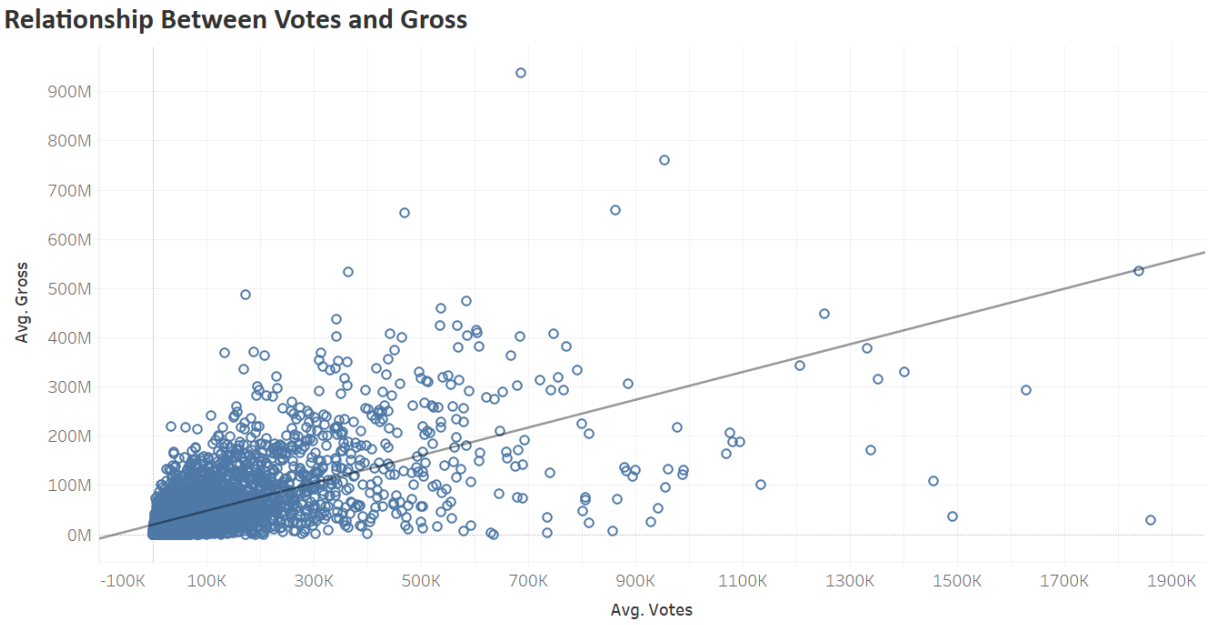
Graph 3. Gross by Score in USD



From Graph 3, we can see movies receiving higher scores are likely to generate higher gross. However, this is not always true. This graph indicates that the movie received the highest revenue is Star Wars: The Force Awakens, which did not receive the highest score. The movie received the highest score is The Shawshank Redemption, but this movie generated very low revenue. While there is a correlation between gross and scores, this positive relationship between the two variables is not obvious. Still, movie studios can refer to IMDb scores to estimate the successfulness of their movies.

As mentioned earlier, the popularity of featured movie stars will also impact a movie's market success. The popularity of a movie star is measured by the number votes the star receives from audiences.

Graph 4. Relationship Between Votes and Gross



The upward sloping trendline indicates that there is a clear positive relationship between votes and gross. We can tell movies generating higher gross are likely to have more popular movie stars featured in the story. We can confirm this by performing a t-test.

Table 6. t-Test

t-Test: Paired Two Sample for Means

	votes	*gross*
Mean	96664.11495	46685719.31
Variance	22540332036	4.42645E+15
Observations	4576	4576
Pearson Correlation	0.641640854	
Hypothesized Mean Difference	0	
df	4575	
t Stat	-47.4381417	
P(T<=t) one-tail	0	
t Critical one-tail	1.645186759	
P(T<=t) two-tail	0	
t Critical two-tail	1.960482649	

The t-Test result confirms our finding.

Summary

The purpose of this analysis is understanding the elements that affect a movie's success in the market. By identifying and analyzing these elements' relationships with gross, the analysis aims to find out ways for movie studios to maximize profits.

The results suggest that there are strong positive correlations between budget and gross and between the featured movie stars' popularity and gross. Animation is the genre that generated the highest gross and G is the most desirable movie rating. There is a positive correlation between IMDb score and gross, but it is not obvious. The results also suggest that the popularity of the featured stars will have clear impact on the movie's successfulness.

There are many ways to improve the quality of this analysis. The data used is retrieved from Kaggle, which was last updated three years ago. It would be better to obtain more recent data for analysis because the numbers may have changed. While IMDb provides open-source datasets, these datasets are very large and have exceeded the maximum capacity of commonly used software programs such as Microsoft Excel. Therefore, a better way to approach this project is to scrap and cleanse the data from scratch and use more advanced technologies such as R and Python libraries to work with the data instead of relying on software programs with pre-written algorithms.