Data Visualizations of Film Revenue

Havan Patel
havangpatel@lewisu.edu
DATA-53000, Summer 2022
Data Visualization
Lewis University

I. Introduction

The film industry is an extremely attractive industry around the globe and film making is a lucrative business. Many well-known film studios have been making films for more than 100 years. Some movies were highly successful and others failed drastically in the box office. In this report, we will be discussing how two complementary datasets can be merged and how this data can be used to find insightful patterns related to the movie industry. We are going to focus on the overall sell of the movies with other attributes and how it contributes to the overall profit.

We have two movie datasets with us which provides information about more than 1000 movies released between 1972 to 2016. Both datasets were downloaded from Kaggle. One of the datasets was "Movie Metadata" [1] and other one was "Highest Hollywood Grossing Movies:[2]. The first dataset contains almost most of the information and attributes needed except the differentiation of revenue made domestically and internationally. That is where the second dataset comes in to help us identify that difference. Also, it contains few attributes but those are very crucial for the report.

An entertainment company or a filmmaker is interested in producing movie, but the decision makers would like to verify certain understandings before investing their money in any movie. The analysis which are available for historical data could provide them with answers in which they are seeking for to better help their decision making. This will also have a benefit to the audience because they will see movies they like to watch while the producers make profit. By doing exploratory analysis it will only help them with their business model. Of course, there are certain factors that contribute to making a great movie like scripts, genre, time, release date and more. We will try to explore several attributes throughout this report to see if there is a significant correlation.

II. Data Processing

With any dataset you use. Doing data processing the most crucial step before doing any analysis. In real world there are so many data and not all data is sometimes needed or always formatted in the correct way. This can skew your analysis and output results that you did not expect. In the first dataset we use we has about 5044 instances and 28 features. The second dataset had 1000 instances and 11 features. If we combine these datasets, we will have 39 attributes and we don't need all of them because some attributes are similar, and some are not useful for our report for example ID attributes. Therefore, we only kept features that we important for the analysis and discarded the others. Data processing was done in python using Jupyter notebook.

There were few steps required for processing the two datasets. First step was that we had to remove some the columns and rename the columns that made sense. The idea here is to remove the confusion and better understand what each column is representing. Second step was to parse the columns that were not formatted correctly. Some of the columns in the dataset contained JSON objects and special characters after the movie tile in which we had to clean. This was important because we wanted to extract each information in the JSON object to help with our exploratory analysis. For example, the JSON object contained list of actors in which we must extract and create new column for each actor. Final Step of this process was to finally merge the two datasets. The merging happens based on the similar columns and it only makes sense to merge them based on similar movie titles.

After the data processing we had 5044 instances and 31 attributes but out of that only 1000 dataset were merged and rest of the 4044 had empty values for missing attributes of second dataset. The attributes increased here because as mentioned above that we had to create separate columns for each of the values in the JSON object. However, not all 31 attributes we used for report, but it is good to keep them just for doing analysis on different values for exploratory reasons.

III. Data Description

After step 2 above, we have following important attributes shown in Table 1. that we will use in our analysis. Some of the attributes were grouped together as it made the most of sense. For example, we grouped the genre category because no one is going to produce film for a single genre of films. There are always at least 2 genres associated with a movie. During the visualization we also ignored null and empty values as it does not provide us with any information. We also further filtered on the data because we did not want to see all the categorical types because then it would be extremely hard to visualize and draw conclusion.

TABLE I. Data Attributes

Attribute	Type	Example Value	Description
Movie Title	String	Avatar	Name of the movie
Genre (Grouped)	String	Action, Adventure	Genre of the movie
Release Date	Date	1990	Which year the movie released in
Gross	Numeric	1204857733	Gross of the movie
Director Name	String	James Cameron	Director Name
Actor 1	String	Johnny Depp	Name of main actor for the movie
Content Rating	String	PG-13	Content the movie is featuring
Domestic Sales (in $)	Numeric	8762937469234	Sales of the movie domestically in dollars
International Sales (in $)	Numeric	21098471203498	Sales of the movie internationally in dollars
World Sales (in $)	Numeric	112324234324234	Sum of the sales of the movie domestic and internationally in dollars
Distributor	String	Twentieth Century Fox	Distributor of the movie.
IMDB Score	String	20	IMDB score from 1- 24
Budget	Numeric	2134324234234	Budget of the movie
Language	String	English	Language the movie was produced in
Country	String	USA	Country the movie was produced in

IV. Methodology and results

Our prediction is that the data will show there is a correlation between the actors being in a film and causing the film to have higher world sales. We also have a theory that the higher rated a film is the more likely the film will make money. We also have the prediction that the budget will correlate to the sales. The higher the budget, the higher the gross earnings. Another prediction is that certain directors have a high correlation with making a large amount of money in the industry. The last one will be that action and adventure genre makes the most money by far than any other genre. Here we are also predicting that family friendly movies or movies rated G or PG will be the highest grossing films. We will test these theories by looking at the data.

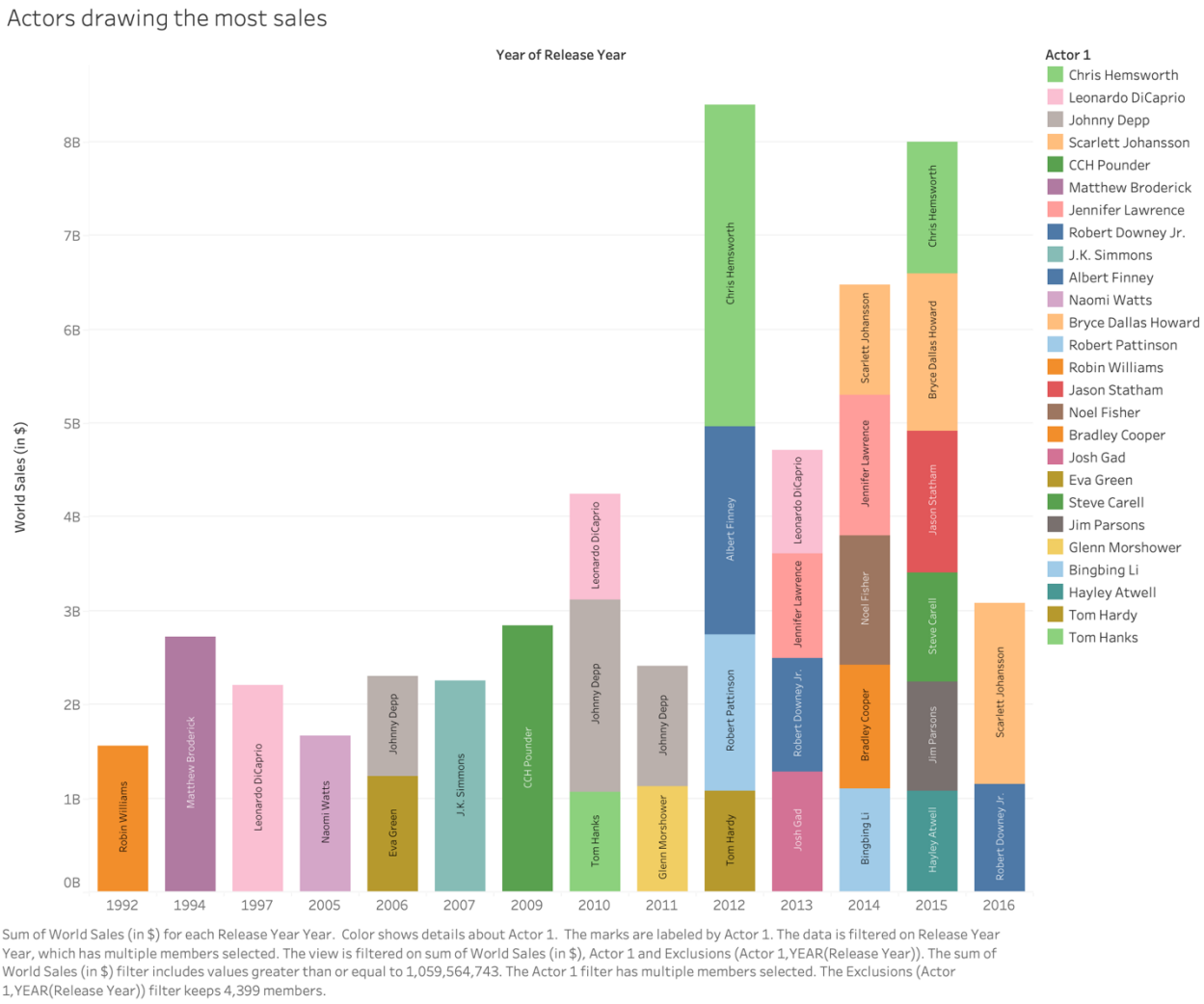

Fig 1. Actors drawing in the most Sales

This is the first set of picturization of data we will be analyzing. As it can be seen there is a correlation between certain actors and the amount of world sales brought in. We can see in year 2012 and 2015, Chris Hemsworth brings in a lot of world sales. In the years 2014 and 2016 Scarlet Johansen brings in a lot of money. We do see a significant correlation between the actors and the movies. One of the highest grossing films is The Avengers. In this film, we can see that the primary actors in the movie are Chris Hemsworth and Scarlett Johannsen. We can also see that the other actors such as Robert Downey Jr, which was also in The Avengers movie was also a significant source of sales from 2012 to 2016.

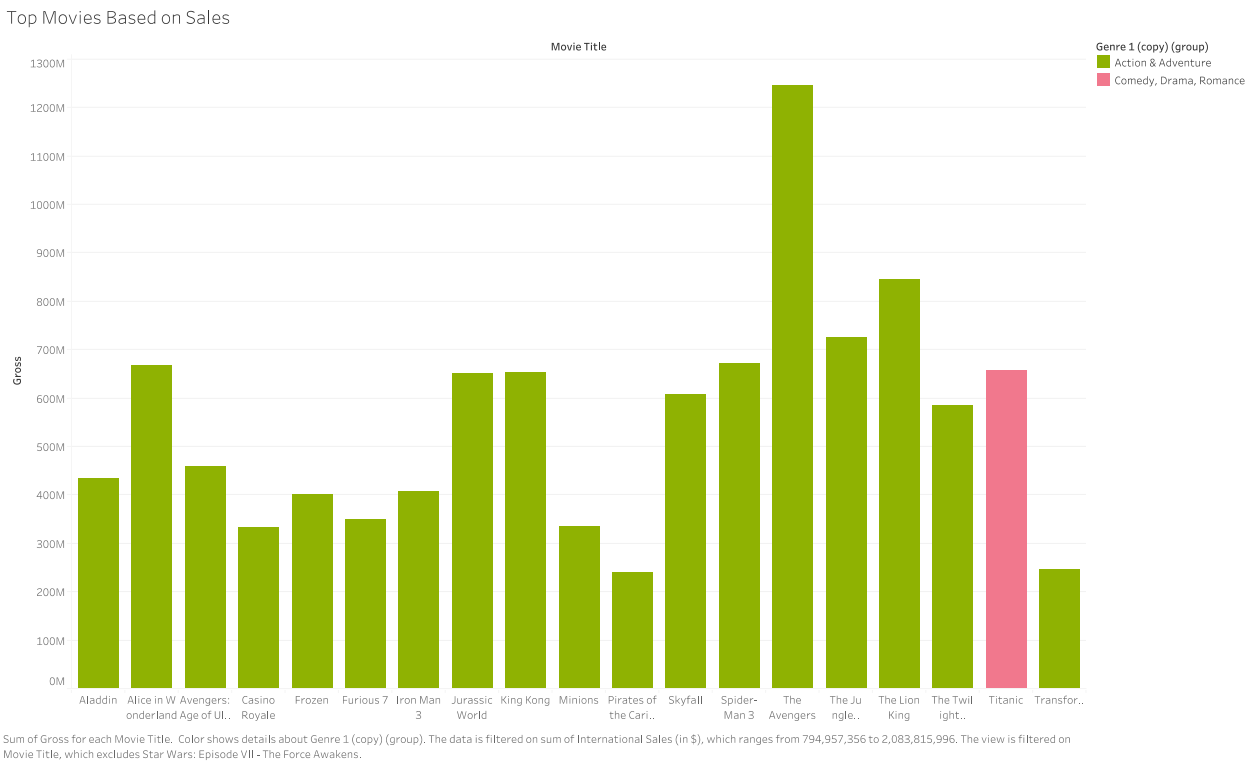

Fig 2. Gross Income for Movies

From this graph this perpetuates the highest grossing films are The Avengers and The Jungle Book. We can see these are some of the highest grossing films. This further perpetuates the idea that certain actors do bring an impact into box office sales. In both of these films Chris Hemsworth and Scarlet Johannsen played significant parts.

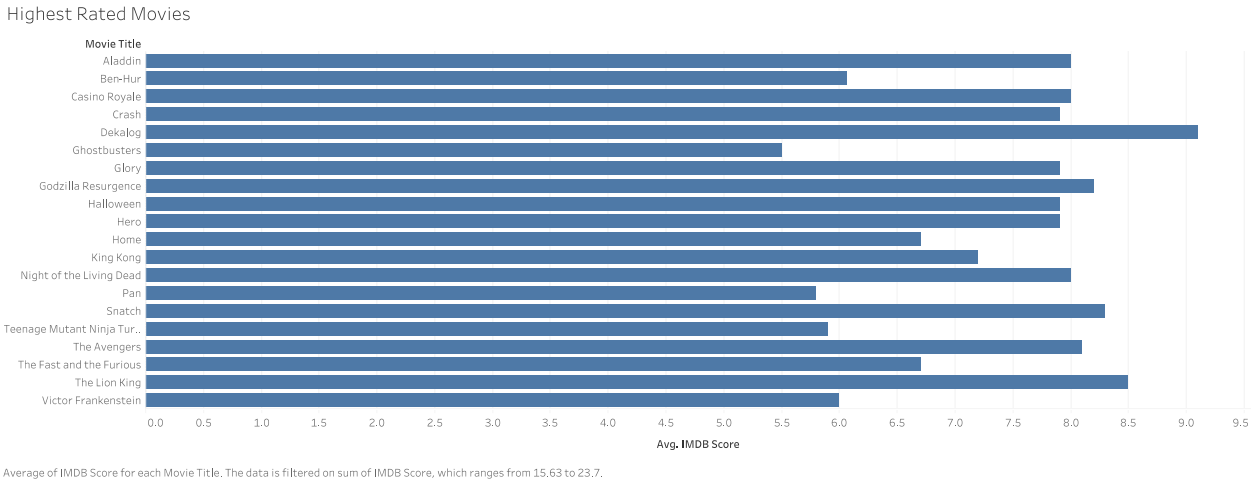

Fig 3. Highest Rated Movies

Here we are testing another hypothesis that highest rated films are the highest box office hits. Here we can see that is not the case from the reviews here. Although there does appear to be some correlation, by the rating alone the highest grossing film should be Dekalog. The second highest grossing film is The Lion King which is one of the highest grossing films and the one right below that is the Teenage Mutant Ninja Turtles. We can safely say there is not a correlation, but there is some impact due to the fact that Lion King, which is the second highest grossing movie the second highest review according to this data.

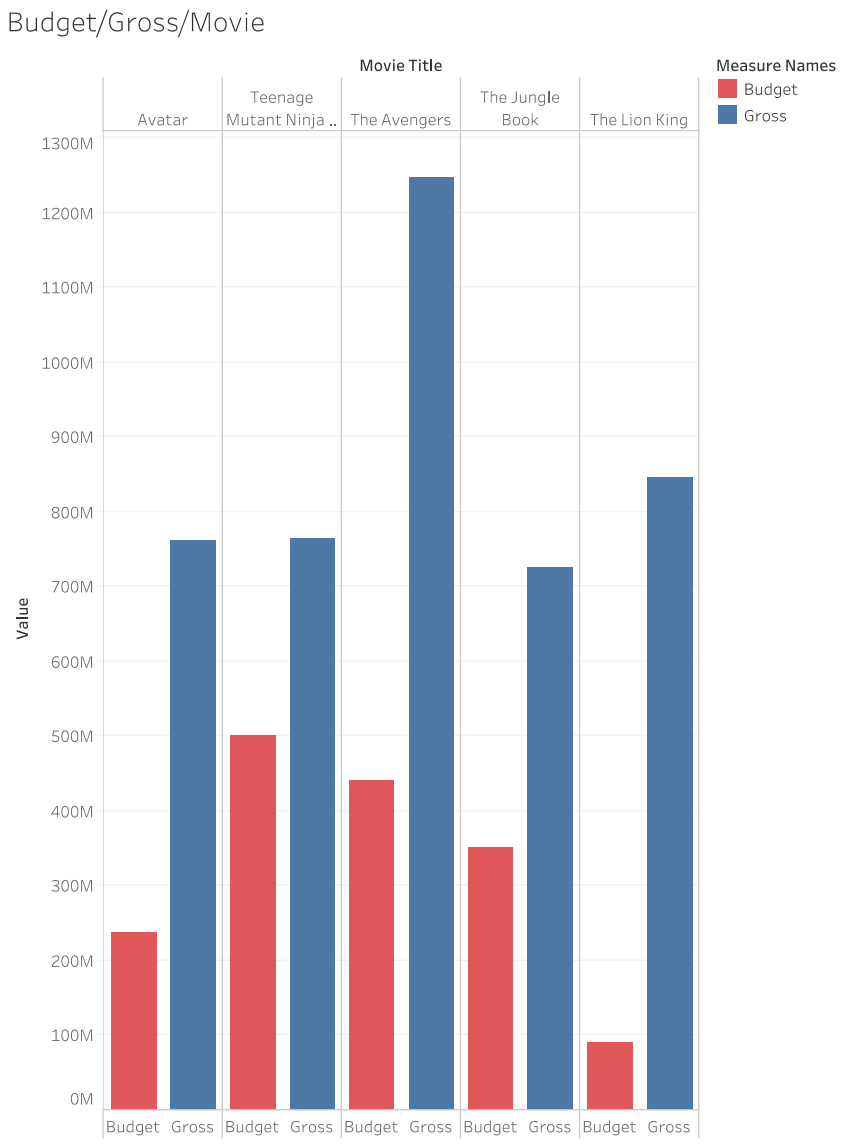

Fig 4. Budget vs Gross

This set of data will tell us whether or not the budget had a large impact on the amount of money the film made. As we can see, the film with the highest budget according to the data is Teenage Mutant Ninja Turtles. While this film was highly rated, the amount of money this film made was not significant. When we look at the Avengers as a film, we can see that even though the budget for this film was lower, this film made far more money relative to the budget. Again, while the budget was low for The Lion King, the amount of money this film made far exceeded the budget. We can safely say that there is not a significant correlation between budget and the gross income of a film. Just because a film has a high budget, this will not make the film a high gross earner.

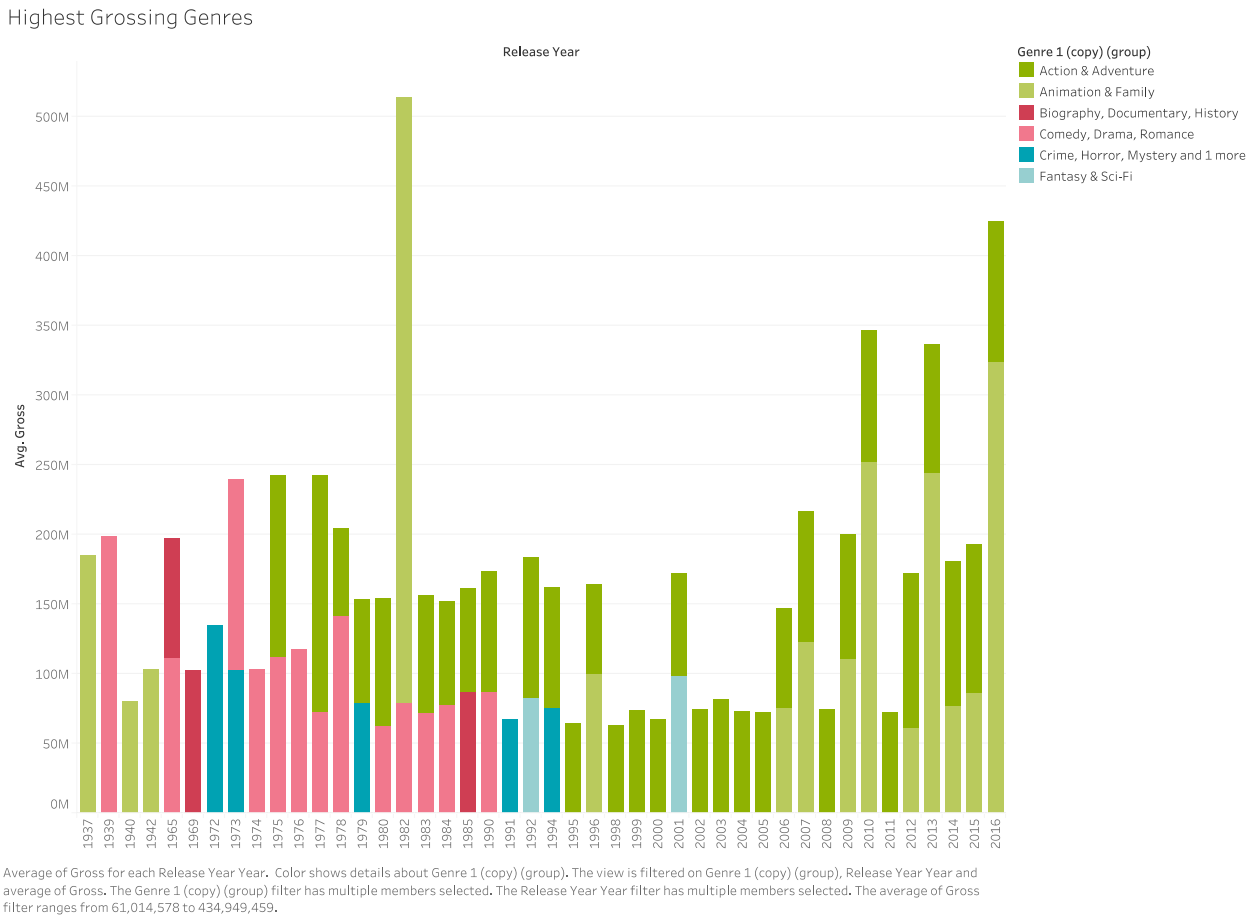

Fig 5. Highest Grossing Genres

As we can see, the Action and Adventure and Animation and Family genres have completely taken over being the highest grossing genres. The Avengers is categorized as Action and Adventure and the same could be said about several other films. The Animation and Family genre is in a close second place and has a significant impact on the film industry as being a very high grosser as well.

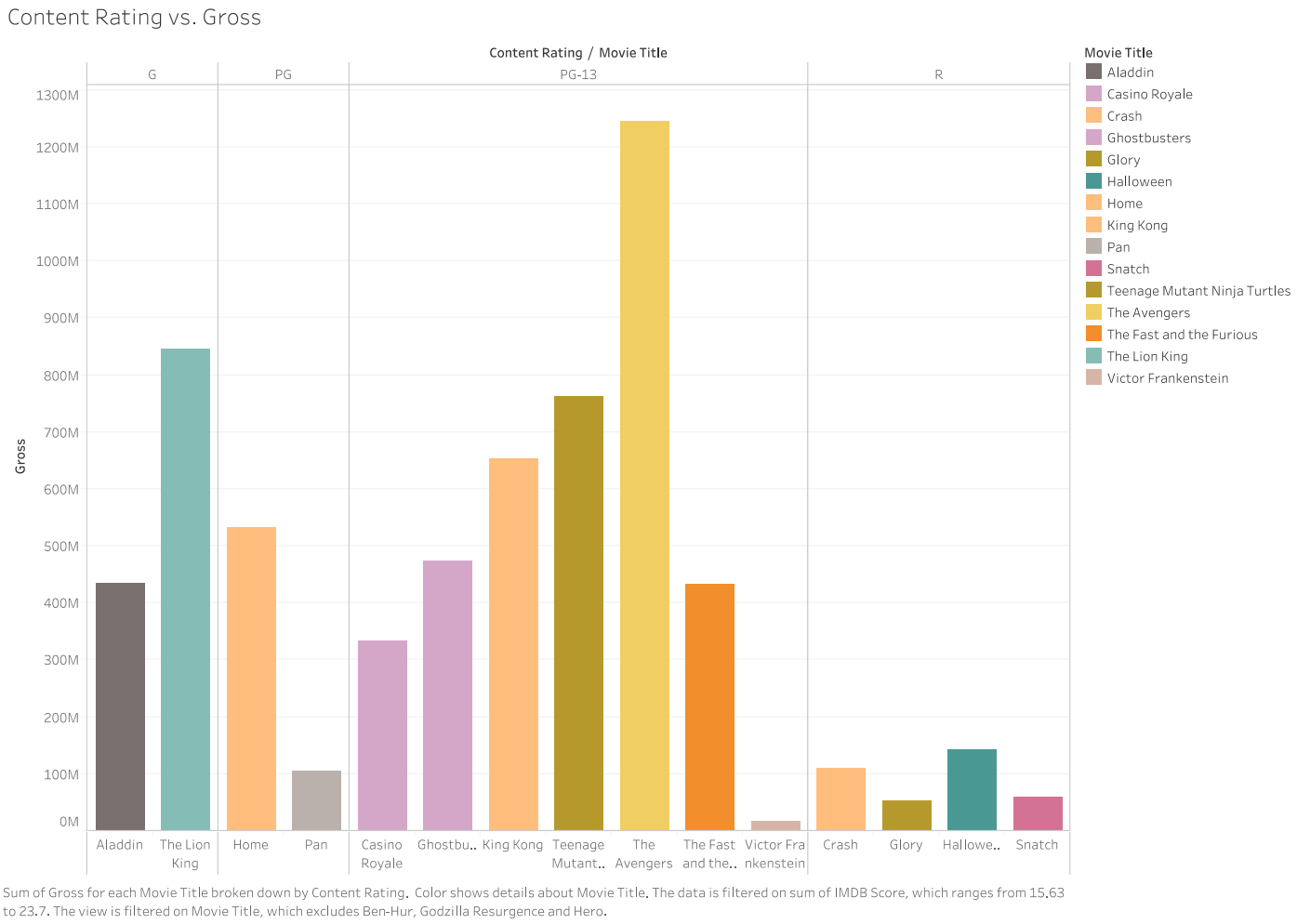

Fig 6. Content Rating vs. Gross Income

We also predicted that rated G and PG movies would make the most money. We found that the PG-13 movies in actuality were the highest grossest movie content rating. Here we see the rating on X-axis and Gross on the Y-axis with the color label done on movies.

V. Discussion

Some of our predictions as discussed did end up being accurate. For example, in Figure 1, the highest grossing actors did show an increase in sales because of their presence. This data is definitive, but there is a correlation. The highest grossing movies according to Figure 2, were the Avengers and the Lion King. There also doesn't appear to be a correlation between highest IMBD scores and movie sales according to Figure 3. According to Figure 4, there is nearly no correlation with budget and gross income. Although there is some relation, this is not consistent to definitive. In Figure 5, we can clearly see the action and adventure genre is the most popular and the highest grossing. We are utilizing Gestalts Laws of color and size. We are using colors to show differentiation between categories. The size clearly shows the significance of the values. The taller the graph, the more impact this has, as the gross amount or world sales stays relatively consistently in the y-axis.

VI. Conclusions

In this exploratory analysis we drew some interesting inferences. We found that actors, and genre of the movie, and content rating does make an impact on the movie sales. The four attributes that we found to be important are: genre, content rating, IMDB rating, actors. If a filmmaker is trying to make a film, they can look at the exploratory data visualizations and analysis to facilitate the movie making process if film makers focus is to make a large gross income. Of course, there are certain attributes which are not measurable like the script of the movie, but we found that the metric we used were impactful metrics to measure by.

REFERENCES

[1] “CHAMBERUNDERGROUND, “Movie metadata,” 08-Jan-2018. [Online]. Available: https://www.kaggle.com/datasets/karrrimba/movie-metadatacsv
[2] “SANJEET SINGH NAIK, “Top 1000 Highest Grossing Movies,” 15-Jan-2022. [Online]. Available: https://www.kaggle.com/datasets/sanjeetsinghnaik/top-1000-highest-grossing-movies