CS 503 – Data Visualization Assignment 2

Yi Ren (002269013), Wentao Lu (002276355)

Our research question

In context of this particular dataset, we are interested in learning what attributes make a movie popular and how are they correlated with each other. In specific, does higher production budget or revenue imply that a movie is more popular? If not, what are the most important factors?

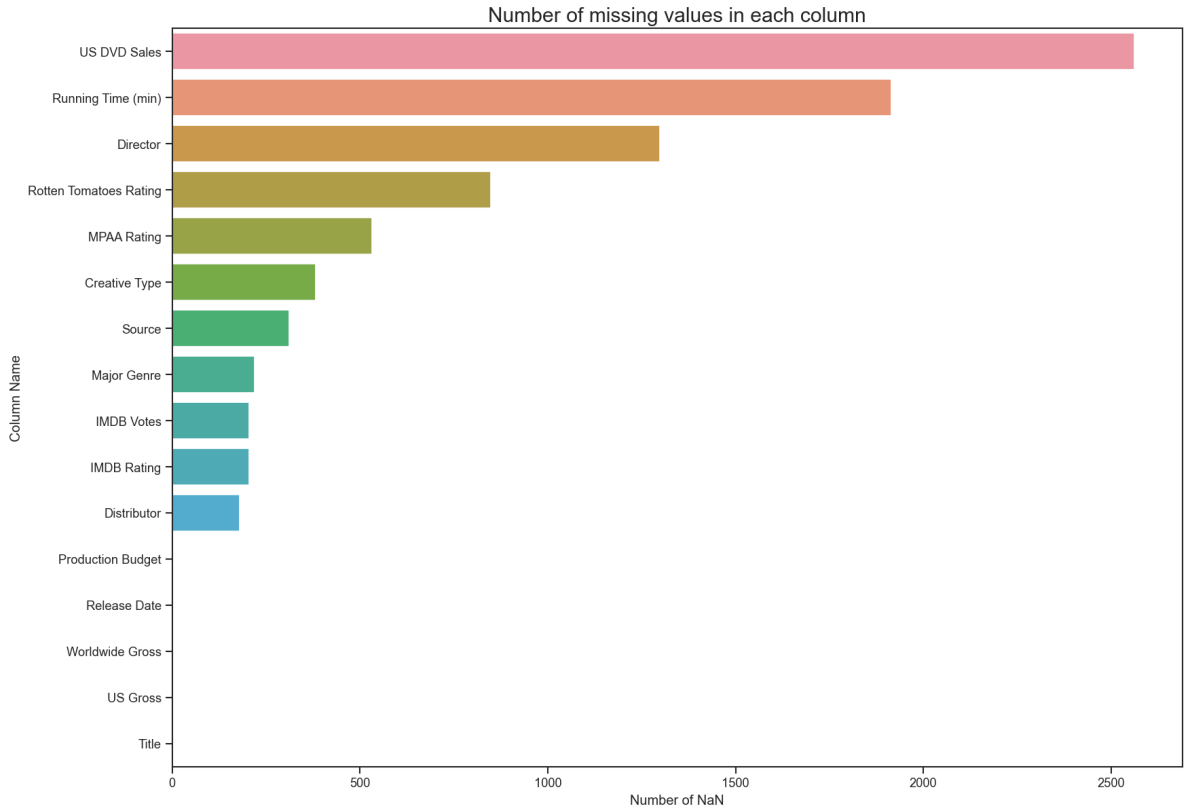
In order to effectively measure the popularity of movies, we are going to use `IMDB Votes` as an estimate. IMDB is one of the largest movies database with millions of users, so the number of votes on IMDB is very indicative of a movie's popularity. In general, the higher the number of votes, the more popular a movie is. On the other hand, the field `IMDB Rating` is more of a reflection of the movie quality rather than popularity, so we are not using it.

It bears mentioning that popularity is not the same as quality of a movie. Some masterpiece movies are bizarre in style and do not meet the tastes of the majority, audience who are interested in such a style tend to rate them high, but those not interested are not likely to watch it. Some movies could be very popular among people who have not even watched them but the quality is just average. Such cases are not rare where the main actors are extremely famous or that movie genre is in its day.

EDA and data preprocessing

In brief, here's what we did to preprocess the data (details are in the Jupyter notebook).

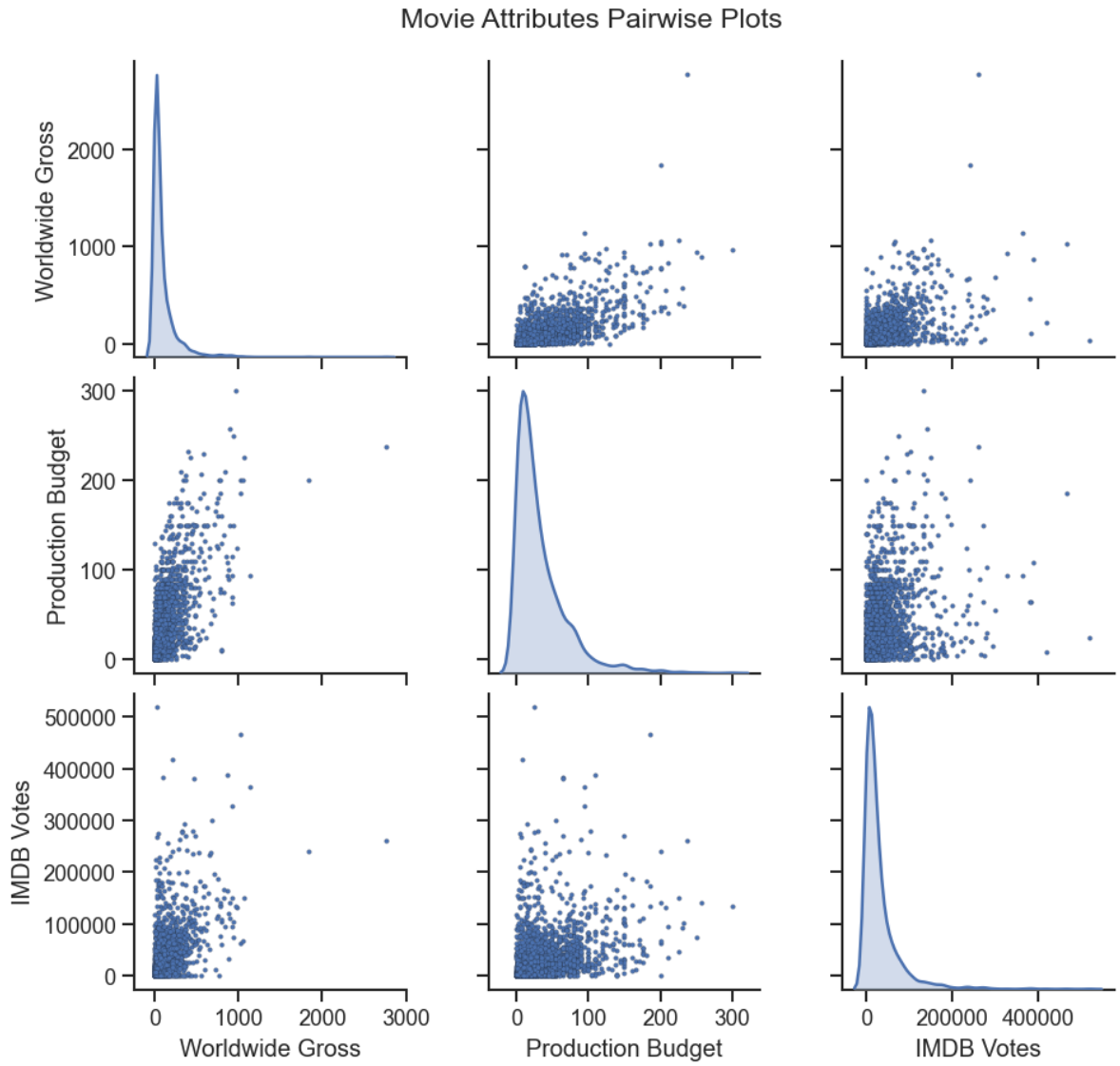
1. Looking at the `dtypes`, we find that `US Gross` and `Worldwide Gross` are strings while they should be numeric, so we convert them into integers.
2. We parsed the `Release Date` column in date format, remove the rows where data is missing (just a few observations). Based on this field, we created a new column `Times`, which divides the dataset into '80s', '90s', '00s' and '10s', even older movies are marked as 'ancient'.
3. We removed the rows where `Major Genre` is missing (just a few).
4. The `US DVD Sales` and `Running Time` columns have over 2000 missing values. Compared to the data size of 3201, this is way too large so we remove them (not useful for our question either).



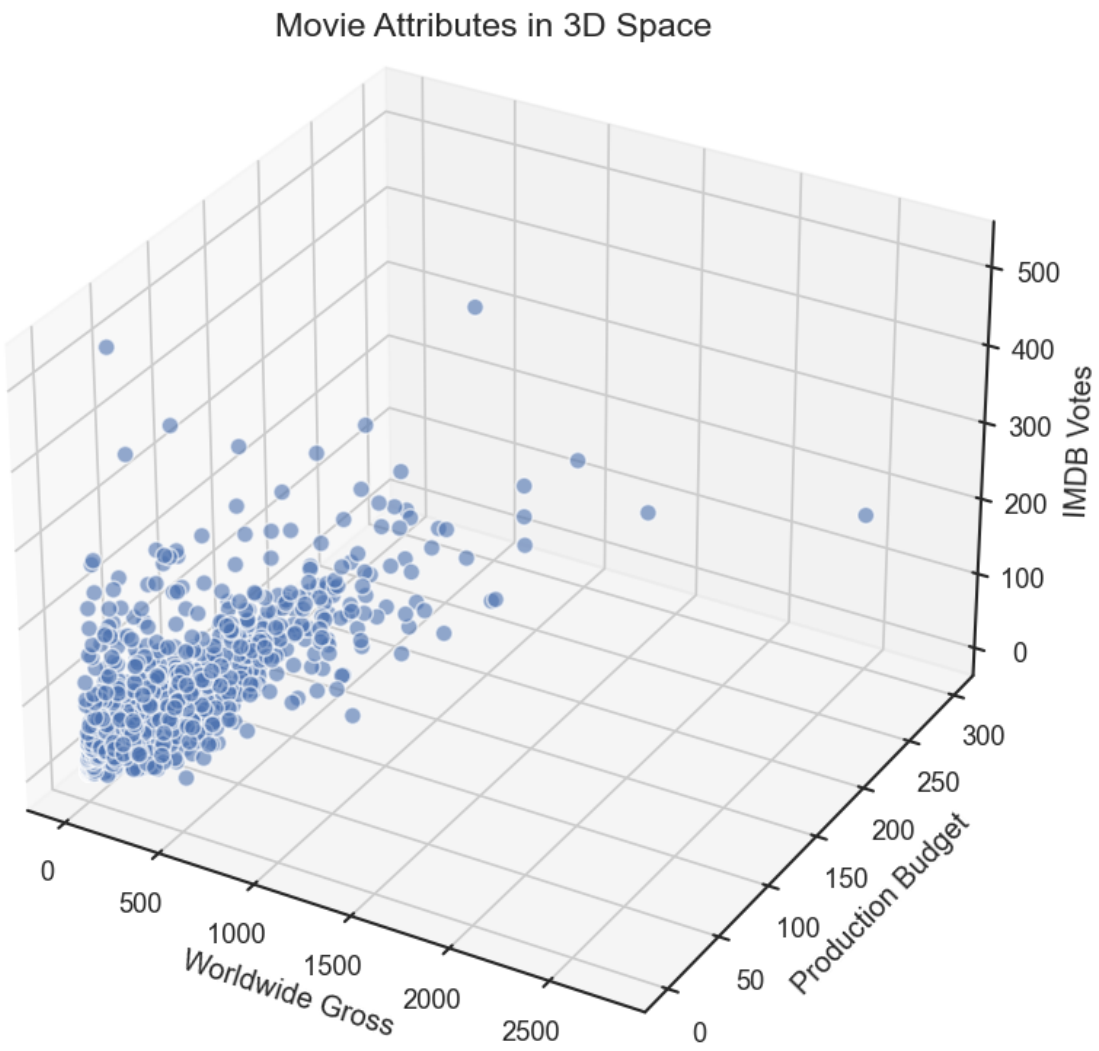
Visualization analysis

For our problem, we are going to mainly focus on these three columns: `['Worldwide Gross', 'Production Budget', 'IMDB Votes']` . To keep things neat, first we converted the units of data such that `'Worldwide Gross'` revenue and `'Production Budget'` are measured in millions of dollars, and `'IMDB Votes'` is in thousands.

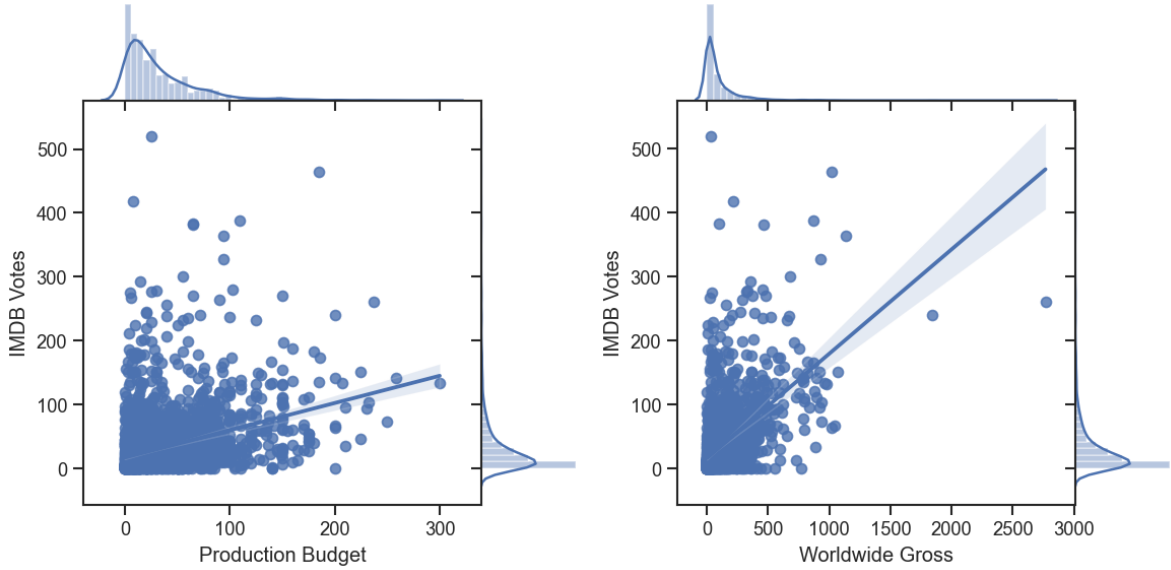
First, we plot the relationship between some numeric attributes. In the pairwise plot below, we can see that most data points cluster around the origin, this means that the revenue, production budget and IMDB votes are not correlated. If we look closer, there seems to be a slight trend between worldwide gross revenue and the production budget, but the correlation is so weak that it is hardly noticeable.



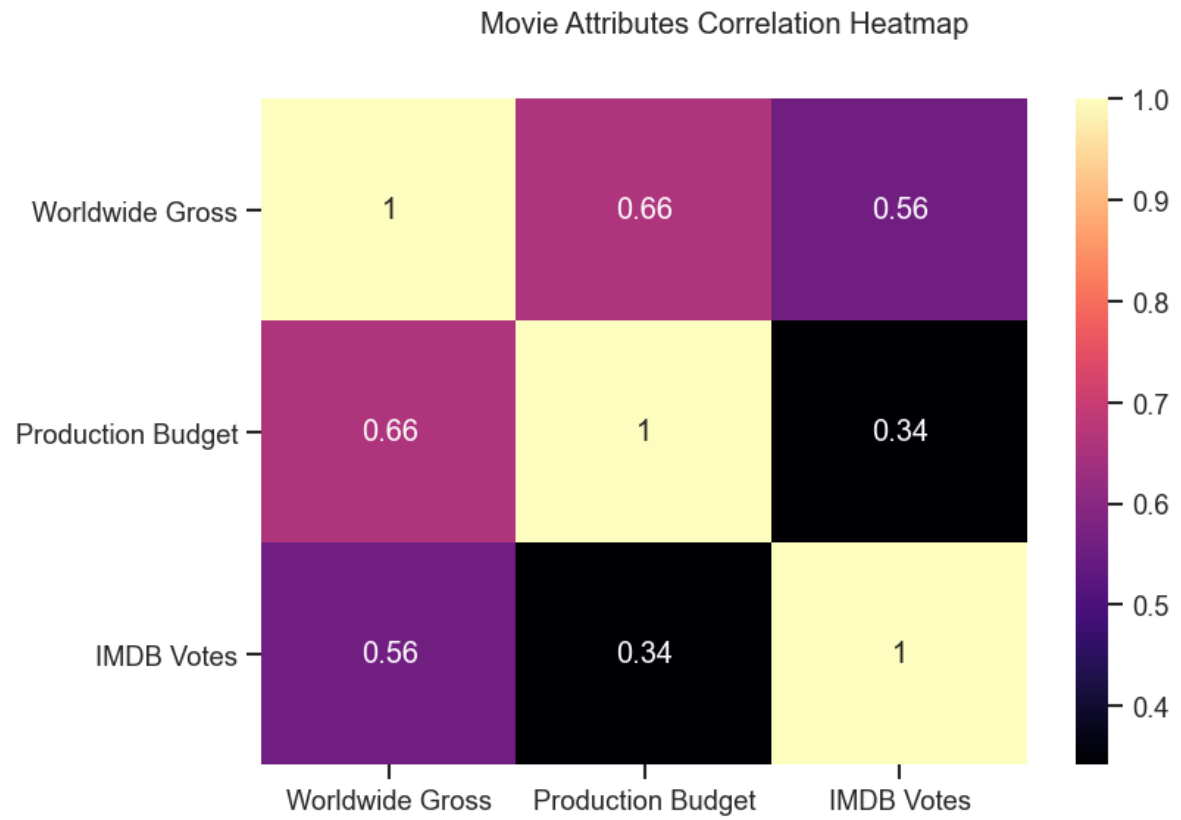
Now we try to plot the same data in 3D space, from the figure below, it's very clear that data points are clustered together without any spatial correlation, this again confirms our belief.



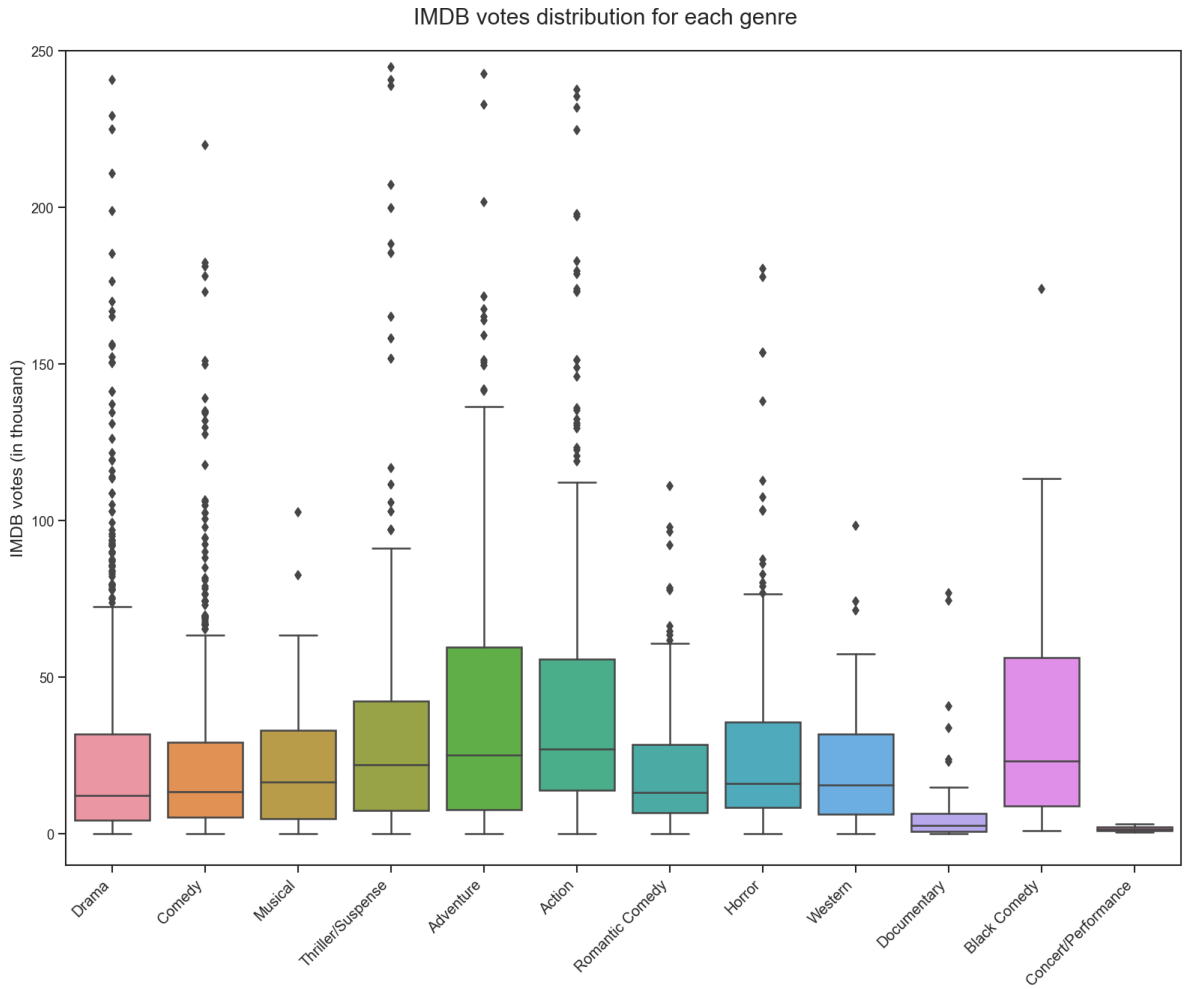
We also visualized the linear regression relationship using a joint plot. Below we can see that the densities of these variables center at the lower left corner, the regression line shows that correlation is not significant, and the number of IMDB votes is quite volatile at any given revenue or budget.



Similarly, a heatmap shows the computed correlation for us. In this case we only see a correlation of 0.6 between revenue and budget, which is still too low to be considered evidence of mutual dependence.

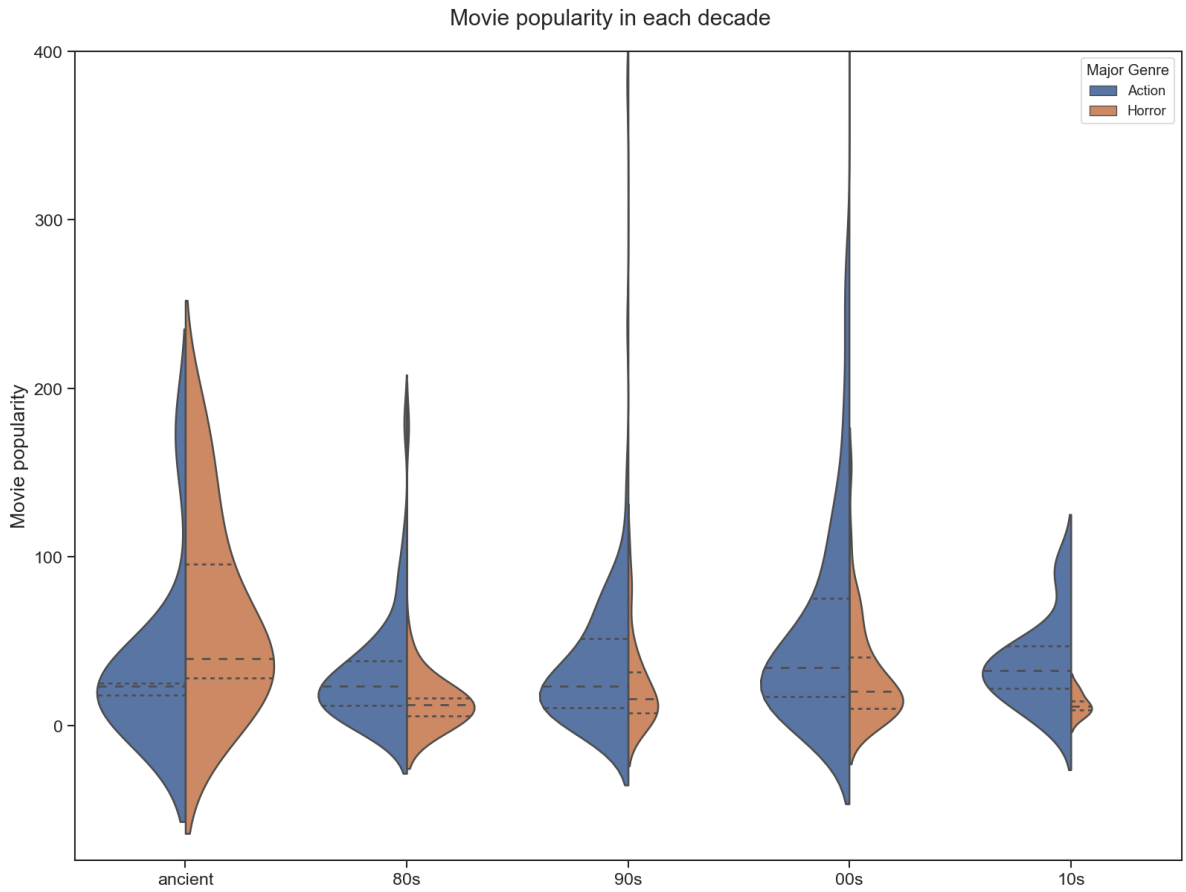


The above result is understandable as movies vary a lot from case to case. Popularity is essentially complex which cannot be easily judged by budget or revenue. Intuitively, some genres can be more popular than others, so let's make a boxplot for each genre.



The boxplot clearly shows that popularity differs from genre to genre, for example, action, adventure and black comedy movies have a much higher median and IQR range than others, but concert/performance movies are very unpopular, its boxplot is almost flat at low values. We can see that drama and comedy movies have many outliers, this means that audience are very diverse in their opinions about these movies.

Movie popularity can also depend heavily on the reputation of main actors, or the history period in which it was released. For example, let's look at how audience voted action and horror movies in each decade.



In the above figure, we see that horror movies are more popular than action movies before 1980, a time period when film technologies are still underdeveloped. Since the 1980s, action movies are becoming more and more popular (the blue median exceeds orange median), perhaps because the technology advances in film have made them much more attractive than before. In contrast, horror movies today tend to lose favor over time.

Conclusion

- In sum, visualization shows that movie popularity can vary a lot between genres and times, but production budget or revenue has very little impact on it.