Project Description

A. Description of the data. Report where you got the data. Describe the variables. If you had to reformat the data or filter it in any way, provide enough details that someone could repeat your results. If you combined multiple datasets, specify how you integrated them. Mention any additional data that you used, such as shape files for maps. Editing is important! You are not required to use every part of the dataset. Selectively choosing a subset can improve usability. Describe any criteria you used for data selection. (10 pts) □

We mainly use one dataset we pulled out from kaggle, IMDB 5000 movie dataset (https://www.kaggle.com/rosendin/imdb-5000-movie-dataset/data). This is a very complicated and comprehensive dataset which contains more than 5000 movie dataset and multiple variables which could be categorized into: basic information (title, genres, durations, year etc.), crew (director and actor/actress), actor/actress social media influence (Facebook likes), financial fact (gross box office and budget) and IMDB rate (IMDB scores and vote numbers). From the dataset, we chose the top 50 highest rating films per decade and put the data into different CSV files(For example, 1970.csv). We read these CSV files using d3.csv function.

B. A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts) □

Our data is to represent the qualities (IMDB rating) and box office performance based on genres over decades. The first modification we did is to group data into decades like 80's present the movies produce between 1980-1989. We filtered out the highest 50 IMDB rated movie in each decade. Choosing top 50 is a tradeoff between the size of data and the data availability. What's more, choosing top 50 per decade means those movies are all highly rated.

In our visualization, we have two portions, the left one represents the quantities of movies in top 50 lists per genre. In other words, it represents the qualities of different genres. The circle which represents that genre will have a larger radius if that genre has more movies in the top list. The right portion represents the box office performance of each genre. The radius of these circles will be proportional to the normalized box share over that decade.

In terms of scales, first of all, we normalize both the quantities and box office performance and represent them in the percentage value(The data of each genre / The data of all genres combined).

After we normalize the dataset, we use these percentages values to map to the radius of each circle. We also use colors in the visualization, each circle represents a different genre and we use different colors to distinguish them. Our original design was to choose only to visualize the top genres and mark other genres as "other". The potential pitfall of this design is the "other" section will always become the largest circle which may cause confusion. Considering the trade between the story of our visualization and the design, we finally choose the color palette recommended by the IMDB and keep all genres.

We also want to show the taste change over decades. This is done by assigning different y values of each graph, so we arrange the graphs vertically, each row has two graphs and they represent the same decade. There is also a year annotation on the left side of the graph. Readers can make comparison between different decades.

C. The story. What does your visualization tell us? What was surprising about it? (5 pts) □
Our visualization can show lots of information when we observe from different perspectives.
1.Horizontally
When we see the graph horizontally, it mainly presents three information:
Which genre has the most high-quality films(from left)
Which genre has the best box office performance(from right)
The relationship between the film quality(rating) and their box office. In other words, if each genre's box office performance correlates to quality.

For example, during the '80s, the drama had the largest quantities of best movies. However, Sci-Fi and Adventure 's box office performance was much better than drama's. In a nutshell, the drama didn't get a box office as good as their ratings while some genres' box office outperformed their ratings.

It is surprising that the situation we found above is quite often, so we are thinking the reason might be good drama movies may have depth but may not be attractive in terms of the technologies and visual simulations. While some genres, like action and adventure, their box office performance are much better than their rating situations. It is reasonable since these genres are top choices if people want to relaxation and entertainment.

2. Vertically
When we see the graph vertically, it mainly shows people's movie taste change over the time.

For example, during the '50s, the drama is dominantly both with qualities and box office performance. As time goes by, the diversity of the film selection increases and it's rarely to see one genre dominants the top list. It's actually surprising to find a trend that the box office of drama decreasing over time.

Another good example is the action movie. It is surprising to know that during the early stage, action movies don't even have a significant place in the top list. However, action movies become one of the most important genres in recent decades. It is not only due to the taste change but also may be the fast development of movie technologies. The same partner also appears when we analyze Sci-Fi genre and Adventure.