IMDB Film Reviews

Team u07
CPCS 4030, Fall 2020

Brandon Mcghee
Sam Newcomer
Dylan Mumm

Introduction

The film industry is incredibly diverse and caters to a wide audience. This ranges from small, indie films to massive blockbusters that gross millions of dollars. We were interested to see what kind of films different moviegoers preferred. We thought it would be interesting to see any relationship between how a film is received by the general public versus the people who take the time to carefully review it. We looked at what genre of film the general public prefer as opposed to what genre of film professional reviewers prefer. We then looked at the change of the popularity of film rating over time, spanning from the beginning of the 1900's until the present.

Dataset

The dataset we chose to use comes from IMDB. IMDB publishes all of the data found on its website. This all comes in the form as multiple tsv files. Included in these files are the listing's title, its rating score, the number of total reviews, the year it came out, actors and directors associated with it, and many others features. This data goes back as far as the 19th century and continues up until the current year. Many of the listings included in the data set were removed. This included any pornographic films, video-games titles, or single television episodes. After these records were taken out we had to condense our dataset by aggregating genres together and taking the total number of reviews from each genre.

Source: https://www.imdb.com/interfaces/

Design solution

In order to tell our story, we decided to use two very different visualizations. Our data consisted of both categorical and numerical data. Because of this, we firstly chose a visualization that encapsulates two dimensions of this data. For our second visualization we took it a step further and introduced a new dimension to our data. These visuals aim to show the contrasts between genre and how they change over the course of the last century.

Our first visualization makes use of a bar chart to depict the difference in both the total number of movies and total number of votes within each genre. Initially, we had these separated into two different visuals but decided to join them together into one with some added functionality. The two had a similar appearance and made the webpage look a little redundant. Combining these two makes the visualization more clean and adds to the level of user interaction. Because our data is both categorical and numerical we made the decision to use a bar chart. Our data was aggregated together based on a shared genre. This represents our categorical data and is displayed on the x axis of our visualization. The numerical data used within our dataset represents what is being measured. In one chart, we display the total number of movies and the other the total number of votes. These are both displayed on the y axis. The marks used within this visualization consist of the line, or bar, drawn vertically. There is one of these for each genre chosen to visualize. The channels used within this visualization consist of the length of the bar drawn and its position within the chart. The length of the bar reflects what we are measuring for each genre. The position of the chart is only used when the sort button is clicked by the user. Once this is done the user introduces a new channel that uses the position of each bar to indicate its length. This new channel adds to the effectiveness of the visualization and improves its interpretability.

The second visualization used implements a streamgraph. This graph still shows the average number of votes within each genre but introduces a new dimension of data: time. This spans from the 1930s up until the present and is measured on the x axis of the graph. The takeaway from this visualization is that the user can see how each genre’s popularity has changed over the course of the last century. The mark used to show the average number of votes is the area that a certain genre takes up. Because the data depicted here is in the form of an area, it was essential to differentiate each genre from another. This introduces our first channel: color. Each color depicts a different genre and how it has changed over time. The other channel used is the size of the area of each genre. We chose to add a level of user interactivity that highlights each genre as the user hovers their mouse over a portion of a specific genre. This also changes the opacity level of the chosen genre in order to make it easier for the user to tell which genre they are focusing on.