Kimberly Brown, Hannah Ajmani, Kevin Ma

Kcb65, hna8, kym5

CS 3300- Project 2

Gender’s Role in Film

**Description of the Data**

1. Overview

For our project, we focused on 763 movies in the last 15 years and how each movie script divides the dialogue between male and female characters. We categorized each film as “male dominated” or “female dominated” based on the percentage of dialogue spoken by male characters (>50% for male dominated, ≤50% for female dominated). We then compared these films to see how many films are male dominated vs. female dominated. We looked at how this division of films changes based on the gender of the crew and the year the film was released. We added data filters for each of these topics. We also used the gross revenue of each film to graph how the revenue changes with the gender division of dialogue. Within each film, we retrieved the names, genders, and word counts of the top credited characters.

1. Description of Database

We pulled our data from [this GitHub source](https://github.com/matthewfdaniels/scripts/) and the International Movie Database (IMDB). The GitHub source provided us with the word counts and genders of every main character in each of these films along with other basic information about each film. IMDB gave us the genders of the first credited writer, producer, and director.

1. Data Selection

Our original dataset came with two parts: a list of movies with IDMB IDs, titles, years, and gross revenues and a list of major characters in each movie with their names, genders, and word counts. With the first list of movies, we scraped the writer, director, and producer of each film from IMDB. We chose not to include any film in which the gender of the crewmember was unidentifiable, thereby avoiding data inaccuracy. We also scraped the movie posters from IMDB (script for IMDB scrape located in project folder). We then filtered the data to only include years released after 2000.

1. Multiple Data Sets

Our datasets included the word counts of every character and the genders of the writer, director, and producer from IDMB. In order to retrieve the percentage of male vs. female dialogue of every film, we used Excel to count the number of words spoken by male and female characters. We then divided these numbers by the total number of words spoken to get the percentage, thereby combining both datasets into one csv.

**Description of Mapping to Visual Elements**

1. Plotting the Data

We have three separate plots in our visual: a Pi chart, histogram, and scatterplot. The first one we created was the Pi chart. The idea behind showing our data this way first was to give the viewer a clear visual of the split between male dominated dialogue in movies and woman dominated dialogue. In order to create this graph, we wrote a function that plotted each data point as a circle in the pi chart, first for all the woman-dominated dialogue films in one area and then for all the man-dominated dialogue films in the other area. There was an issue with the sizing of the pi chart circle, which was correlated with the number of data points being graphed. We did our best to minimize the size of this circle while fitting all the data points, but the results weren’t exact. The color of each data point was designated as either red or blue depending on which gender dominated the dialogue in the film. This way both color and placement in the Pi chart differentiated male dominated films from female dominated ones.

The histogram was plotted in order to show a further breakup in the percentage of dialogue. We kept the color of each data point the same as in the Pi chart to maintain consistency across our visual. The x-axis is the percentage of male dialogue in the film and was scaled linearly from 0 to 100 percent (with 21 bins total) while the y-axis is the number of films and was scaled linearly from zero to the number of films in the largest bin plus five. Our idea for the histogram was inspired by [this breakdown of dialogue](http://polygraph.cool/films/). However, we only used this site for visual inspiration and used the histogram codes outlined in class and homework for our project.

The scatterplot was our last visual and was created in order to show why most scripts appear to be dominated by men. Once again we kept the color of each data point the same. The x-axis is the percentage of male dialogue in the film and was scaled linearly from 0 to 100 percent, while the y-axis was scaled linearly from 0 to the highest revenue a film made in the dataset. The size of the circles is another representation of the gross revenue of each film in comparison to the other data points. We calculated the radius as a multiple of the square root of the film’s revenue divided by the highest revenue in the dataset. By changing the size of the circles, we give another visualization of how much revenue a film made aside from the location along the y-axis.

1. Filtering Data

In order to give the viewer more detail about the dataset, we allow the data to be filtered by the gender of the producer, director, and writer as well as the year the film was released. These filters can be applied across all three visuals and give a clearer view of the data especially in the last graph. All four filters can be applied together or separately. When applied in the Pi chart, the size of the graph is reduced and a grey circle is created to show to size of the graph when no filters are applied. The viewer can get a general sense of what portion of all films meet the filter requirements. In the histogram and scatterplot, the filters can affect the scale of the y-axis depending on which data points are removed.

1. Adding the Popups

The popups allow the viewer to determine which dots on the page correspond to which films. For example, if a user wants to see which dot had the largest revenue of all films, they can scroll over the dot with the largest area and see that it was Avatar, released in 2009 with $882M in gross revenue. The viewer can also see that 66% of Avatar’s dialogue is spoken by male characters and that all major crewmembers (except for the producer) are male. To enhance the aesthetics of the popup, we also added the movie poster to each film’s popup information.

The position of the popup follows the position of the cursor. Within the popup, the bar graph colors correspond to the gender of dialogue; the bar representing male dialogue is blue and the bar representing female dialogue is red. We also added color to the genders of the crewmembers to allow the user to easily determine whether the crew is male-dominated or female-dominated.

**The Story**

Our visualization shows how male dominated the film industry is in Hollywood. Not only are the majority of characters dominated by men, but also so are the creative roles behind the camera. The absence of women behind the camera can be seen in the lack of data points when male directors, producers, or writers are filtered out. However, the visualization shows that when women are a part of the creative process behind the camera, there is either a more even distribution of speaking roles or women dominate the dialogue. For example, displaying films that are written by women increases the number of female-dominated films by over 40%. In this way the lack of women behind the camera is one explanation of the lack of dialogue spoken by women on camera.

One thing to note is that changing the gender of the producer does not drastically affect the percentage of female dominated films. This is because typically a producer is tied to a production company, and although the producer does have some say in which films they produce, they must first receive the approval of the company. In this way, a female producer is more likely to produce a film that aligns closely with the other projects of her company than one with heavier female dialogue. An interesting further analysis would be to examine the production companies of each film and see if certain companies lean more towards male dominated films than others.

While the Pi chart shows a binary display of gender-divided films, the percentage histogram expands this display to show just how skewed these films are towards male dialogue. From looking only at the Pi chart, it is easy for the viewer to question if this binary scale is an accurate depiction of gender dialogue. While 20% of all films are female dominated, there are several films where female dialogue is 49% of the total, and the viewer can question whether it is fair to place those films in the “male dominated” category. To address this issue, we display this histogram to show that in fact, most films fall within the 80-90% male dialogue category. This demonstrates that not only are more films male dominated than female dominated, but the majority of films heavily focus on male dialogue.

The revenue scatter plot answers the question of why films are so skewed towards male dialogue. This plot compares the division of dialogue to the film’s gross revenue, and from this plot, it is clear that films with a higher percentage of male dialogue receive higher levels of revenue. This is key to understanding why films are more likely to have more male dialogue. For some time now, there have been huge debates within Hollywood pertaining to gender and racial discrimination within the industry. While some say the awards ceremonies are biased, others believe the root of the problem is within the casting. However, this plot shows that male-centric films are more appealing to audiences and therefore produce higher revenue levels. Instead of criticizing the industry for producing only male dominated films, we should recognize that Hollywood’s audience is more inclined to enjoy these films, and the producers are only doing their jobs to maximize their revenue. The solution to the gender bias is not only to increase the female population in Hollywood, but also create films that are female-centric and still appealing to audiences.