McKenna Quam – DS 4200 HW 5

Dataset: TMDB 5000 Movie Dataset

For this assignment I wanted to do something with movies because they are something that I enjoy and there are a lot of datasets out there to do with movies. I found a data set on Kaggle that had the top 5,000 movies from the Movie Database, a website that catalogs movie as well as their casts. The website also allows the users to rate films. I just used the information about movies in this assignment as having the entire cast and crew for each movie was more detail than I needed for this assignment.

Possible limitation of this data set is that there are less entries for films before the 1990s. This is due both to the fact that there are less films that were made the further you go back in time but also because older films are less popular with a modern audience. This means that this data set is not ideal for temporal analysis as due to how the films were selected, as there is a bias towards more modern films.

I want to note that the movie database calls it's user reviews "votes" I know that some other cites call them ratings so I wanted to be clear that when I say votes in this document or in the visualization I am referring to the amount of user reviews.

Visualization 1: Budget vs Revenue

For my first visualization I wanted to explore the correlation between the budget of a film and its revenue. So, I created a scatterplot with the x axis showing the budget and the y axis showing the revenue. When I did this, I noticed that it wasn't easy to tell which films made their budgets back because the graph wasn't square and because the numbers on the axis were very large and therefor harder to compare. I did 2 things to fix this: first I created two new derived columns that had budget in millions and revenue in millions. This reduces the number of zeros a person is reading when looking at the axis and therefore makes it easier to understand. Second, I added a conditional color to the dots where is the revenue was greater than the budget the point was colored green and if it was less than it was colored gray. This means that the user doesn't have to read the axes at all to know if a film made money. I choose green to represent a profit as green is associated with money and grey to represent the losses because while black is usually associated with loss, I felt black was too distracting from the green and thus grey was a better color.

Next, I wanted to allow the user to filter the data within the chart as 5000 points on a scatter plot is a lot to look at. I decided I wanted the user to be able to filter by a film's popularity, as that would allow for exploration if more popular films made more money. I did this by creating a tick graph represent each movies number of votes to display beneath the

scatter plot. I chose a tick graph because not only does it allow the user to filter data in the revenue chart, but because it is a frequency chart itself the visualization now communicates how many movies have 1000-2000 votes vs than how many have 6000-7000 votes. From playing around with the filter I was able to surmise that less popular films are much more likely to lose money, which is common sense however I could see some outlier gray dots when I was filtering.

To provide detail to those outliers I decided to make a mouseover tool tip that displays the title, release date, and number of votes a film got. This allows for the user to investigate induvial films on the graph. For instance, one of the gray dots I saw with a high number of votes was Donnie Darko, a horror film that on its release in 2001 didn't do well at the box office but as time as passed become a cult classic hence how it lost money but got a lot of votes. This provides color to the data, rather than being just a bunch of dots they are individual films. It is also just fun! As I was tweaking the visualization, I saw Shrek 2 pop up which is film I really enjoy so it made me happy to see it. I chose to display title because that is how most people identity a film, the release date because some films have the same title, as well as the number of votes because while the user knows what range the votes are in, they can't know the exact count without the tool tip.

Visualization 2: Revenue and Count of Films over Time

For my next visualization I wanted to show the different trend in what kinds of films were produced over time. The dataset I chose has a column named genre which holds each of the film's genres in a map with the key being that genre's id and the value being the name of the genre, however when the data set was read from the csv the map was made into a string and so I had to process it into a useable form for analysis. I decided to take only the first genre from each map as for my visualization I only wanted to count each film once. I wanted to show both how many films of each genre were being made as well as how much money they were making respectively so the top chart shows the sum of the revenue of all films of that year by genre and the bottom graph shows the number of films that were made in that year by genre as an area graph. There were a lot of different genres of film so I filtered the data set to only include the top 10, but that is still a lot of categorical variables to color encode so I also created an interactive legend so the user can select which genres of film they want to be viewing specifically. This allows for easy comparison between genres when selected. For instance, I can see that the two most popular genres of comedy and drama have had a similar number of movies made of each every year, but that in 2010 comedies made way more money than dramas.

Sources

Movie database: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
Altair documentation: https://altair-viz.github.io/index.html
Altair examples: https://altair-viz.github.io/gallery/index.html