5609 A2

Why do we use opacity = 0.5 for points in the scatter plot? What would happen if the default transparency of all points were set to 1?

We use partial opacity = 0.5 so we can see when points overlap. With fully filled in points (opacity = 1), you'd only see the top point when multiple points stacked on top of each other, which hides information about how many points are in each area. The opacity = 0.5 also helps the selected point stand out more since it becomes fully opaque when clicked. When you have hundreds of movies in your dataset this really helps the points become more visible.

In the current implementation, all attributes can be bound with the three visual channels, leading to inappropriate encodings (e.g., using the x-axis for the title of 895 movies). Which attribute should be removed from the dropdown list for each channel?

In the current setup several attributes should be removed from the dropdown lists to prevent inappropriate encodings. For the X and Y text fields like "primary_title," "original_title," and "simple_title" should be removed because they contain too many different unique values (899 different movie titles) that get squeezed onto an axis. The "tconst" ID field should also be removed as it's just a random identifier with no meaningful or real order and "genres" creates confusion by generating multiple points for the same movie. For the size channel, only number fields that truly represent quantity should be included: "runtime_minutes," "average_rating," and "num_votes." Date fields like "year" should be removed from the size channel options because time values don't represent size for example a 2010 movie isn't a size but rather a time.

I use curves for the year distribution to showcase how to implement the path generator. But is using curve the best choice here? Anything wrong with using the basic curve?

I see how curves are used in the example. I created both a line chart and a general curved chart but with a basic curve it’s hard to see where each curve in the graph is. Rather than hitting each data point a general curve can misrepresent large amounts of data because it evens out the curve.

First, please list three potential factors that might influence the rating of summer movies based on your prior knowledge or intuition.

The 3 I would list are Genre, Number of Votes and Runtime

Then, use the multi-view visualization to investigate the influence of the three proposed factors.

For testing Genre vs. Rating:

X-axis: "genres"

Y-axis: "average_rating"

Size: "num_votes"

For testing Number of Votes vs. Rating:

X-axis: "num_votes"

Y-axis: "average_rating"

Size: "runtime_minutes"

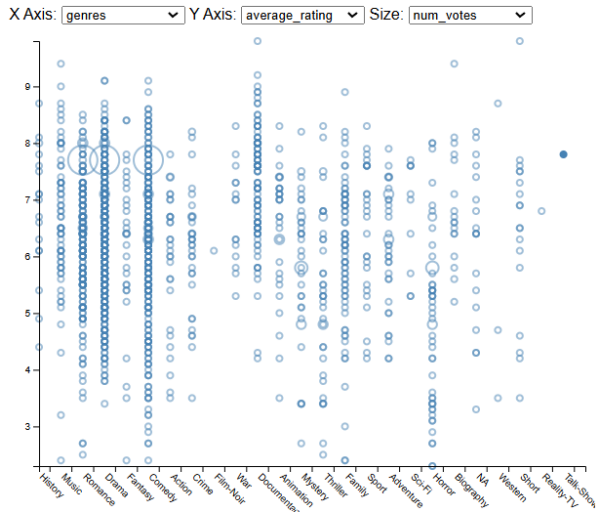
For testing Runtime vs. Rating:

X-axis: "runtime_minutes"

Y-axis: "average_rating"

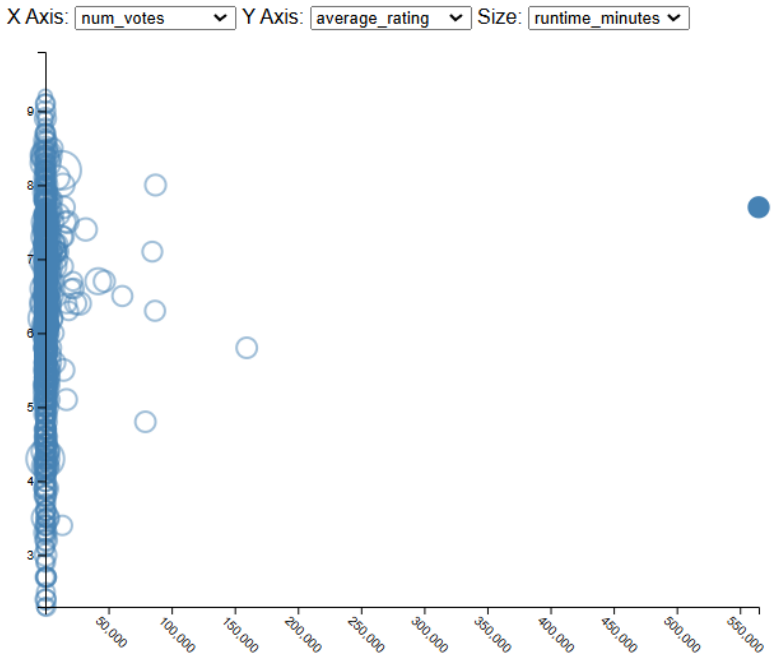
Size: "num_votes"

Looking at the first chart with genres on the x-axis and ratings on the y-axis, I can see some clear differences between movie types. Documentaries, biographies, Music and history films usually get higher ratings if not the highest - many of them score above 7.5. On the flip side, horror and action movies tend to score lower overall with the highest action rating at only a 7.8. I noticed that dramas, comedies, and action films have bigger circles meaning more votes, so they're more popular even if they don't always get the best scores. It's interesting that some smaller genres like Talk-Show have fewer movies but they tend to be rated pretty well.



The second visualization shows votes on the x-axis and ratings on the y-axis, and there's a clear pattern here. Most movies have relatively few votes, creating that dense cluster on the left side. But when you look at the super popular movies like the ones farther to the right with tons of votes, they almost never have terrible ratings. It seems like there's a connection between how many people

watch a movie and its quality - the more votes a movie gets, the less likely it is to have a really low score. The bigger circles represent longer movies, and these tend to appear more in the higher rating areas.



In the third chart comparing runtime to ratings, there's a less obvious but still noticeable trend. There's a slight upward pattern suggesting longer movies tend to get somewhat better ratings. The very short movies (on the left side) have ratings all over the place, while movies between 90-120 minutes make up the majority of data points and cover the full range of ratings. The longest movies rarely get terrible scores most maintain at least decent ratings. The bigger circles appear more often in the middle-to-upper rating range, supporting what we saw in the second chart about popular movies typically having decent ratings.