Michael Daum
CMSE 402-001 Spring 2024
Honors Writeup

What was your motivation for choosing your visualization tool?

I chose Power BI for a few reasons: I was really interested in involving interactivity into visualizations, Microsoft office products are frequently used in workplaces making it beneficial to be familiar with many of their services, and I would have liked to use Tableau more throughout the semester but I never felt I had a good opportunity to make use of it.

How does the project you developed draw on your knowledge from class?

The project employs the shaded area graphs that Tufte mentions on page 187, using no additional effects to highlight the differences in the dataset. Additionally for the tree map values are displayed on the graph since in the elementary perceptual hierarchy comparison of area tends to be more difficult. All of the graphs employ a similar color scheme with movies and TV colored differently for everything but total rating count in the tree map, however that uses a single shade gradient of one of the other colors used. Furthermore, with the interactivity this visualization lets the user explore more with a minimal amount of guidance in a drill down sort of interactive visual, containing some elements of a guided slideshow (encouraging the viewer to select either TV Series/Movie) while still letting the viewer explore rating elements. In addition, noting some of Cairo's suggestions regarding missing data, I chose to add each data point as a dot in the shaded area graph so that it was clear that the lines did not represent data that wasn't there. This is also the reason that the first plot is a scatter plot, since it wasn't very clear where there were missing segments of data at first glance, misleading the viewer. Additionally, Cairo mentioned using a log scale for fast growing data sets, which the top left scatter plot exhibited a clear trend of massive growth over a small time frame, leading me to implement this scaling. I also used detailed axis labels and titles so that viewers are aware of what the exact quantities they are viewing are.

What struggles did you run into when developing this project?

The dataset wasn't as clean as I thought it would be, with importing it immediately leading to an error. This helped me to quickly pick up the filtering options that were available in Power BI, removing the extra work I was thinking I would have to do in python. Another issue I had was that the restrictions of some of the tools were quite limited, such as putting multiple columns of y-data on the same scatter plot. Additionally I had a column of integers that wouldn't plot until I went and changed the data type inside of the data source. Other problems included it automatically trying to summarize data, not using intuitive manners to achieve certain results (can't scatter multiple metrics on the same plot), and unfamiliarity with Excel which would be very helpful in cleaning the data and creating new columns.

What part of the project are you most proud of?

While it might not seem like a lot, the filled area graph took me quite a bit of time since I had to directly work with filtering the source and creating new columns to do so. Since Power BI is by Microsoft, there were various Excel functions that I'm not familiar with that were available. Ultimately I had to spend a lot of time trying to remove the units from the durations while trying to keep the data separate. Ultimately I decided to leave it combined and just filter by TV series or movie so that the axes units aren't mixed. This is also the primary purpose that a filter option is available in the middle of the report, since it encourages removing the dual unit axis since all of the data in that column are integers that aren't separated by their different units.

If you had more time, what would you change or do differently?

I really wanted to try and split apart the genres and create a plot that contained information regarding genre counts per year that were added to netflix, however since multiple genres were all in one column I would've needed to create multiple columns splitting apart every ',' delimiter and then plot all of those together which wasn't unreasonable but I didn't have the time to do. Additionally I would've liked to make the tree map of rating counts per year more interesting on its own, since it doesn't really provide much except for when a rating is selected showing how much of the grouping is made up of that particular rating. A different dataset could have also drawn out more of the tool, especially given the integration with excel features allowing combinations of different sources. Additionally finding some way to highlight the ability to drill-down in that slide through the ratings would be interesting and add quite a bit to this visualization in my opinion.