Phillip Thomas

DAAN 871

Dashboard Purpose

Movies are common subjects of explanatory and exploratory dashboards on the internet, and it’s not hard to see why. As primarily commercial projects, movies are metric-rich by nature. As such, their quality has been compared, even gamified using many attributes: revenue, development costs, number and type of award nominations, genres, directors, and that most controversial of metrics, ratings. From the beginning, the project had been personally motivated by an interest in observing, by using data, the phenomenon of public opinions on movies shifting over time. The main question(s) to address naturally followed: which films experienced the largest shifts in public opinion from the time they released to the present? Which ones remain as highly or lowly-regarded today as they did back then? Are there any obvious examples of this happening in popular culture that are likewise reflected in the data, such as a film widely known for being polarizing?

Background Information

While the fact that these questions were being raised was a good sign of the project’s potential insights, many of the datasets available were made with summary statistics in mind, such as an aggregate IMDB rating at the time of collection. Clearly, it wasn’t quite as easy to gather timestamped data for that purpose. This limitation led to be selection of a large dataset from Kaggle, the “20M MovieLens dataset”. It contained the time series data needed, including the year down to the second when a given user submitted their 0-5 scale rating. Two separate csv files in the dataset, rating.csv and movie.csv, were joined into a table in a new file titled “Final movies.csv”.One limitation of this data that was immediately apparent upon testing was that not every film had been rated for the same amount of time or in a consistent fashion, leaving gaps and uneven points and lengths in some time-series data. Additionally, the overall timespan ranged from January 1995 to March 2015, so no movies released after that month were included. Even then, with thousands of movies represented, it was clear that an ordinary user would likely find something to interest them, given some tools.

Users

To clarify what an “ordinary user” represents, the dashboard was intended to be used in a casual fashion by an average movie viewer with little if any professional experience in moviemaking or criticism. Sites like Letterboxd, which are intended for enthusiasts and amateur (as well as professional) critics are known for being accessible and providing aggregate statistics, such as user ratings for every film, but not all the common metrics discussed in my first paragraph are available. At the very least, they are separated from the main statistics to maintain focus on a handful of metrics at a time This simple-yet-deep approach informed the type of casual yet interested audience I was going for, not market analysts or finance professionals.

Data and Visualizations

First visualization

The initial vision for the dashboard first centered around a line chart that could take advantage of the timestamps in the data to visually demonstrate peaks, valleys, and overall trends in ratings for each movie. Since it would be difficult to represent thousands of films, it was decided that the time series would employ at least one filter by which the display could be controlled to show one or a handful of films of interest at a time. Important attributes included the title (including the year of release in parentheses), timestamps, and ratings. Tying back into the simple approach I mentioned, the granularity of the timestamps was reduced, showing average ratings only on a yearly basis for the purpose of long-term trend analysis. Because each film's genre tags for the data were contained in one column and were cumbersome to create filters and parameters with after splitting, it was decided that filtering on that attribute would be reserved for the second visualization.

Second visualization

This second visual would take that time series information and convert it to a sorted bar chart to allow users to see which movies experienced the greatest long-term increases and decreases in ratings for their respective genres which could not be as easily discerned from a crowded line chart, even with tooltips. Even though a non-condensed bar chart could offer a view of the films that weren't closest to the extremes of rating changes, I retained only the information that viewers would be most interested in, mainly the top and bottom 'X' number of films that experienced the greatest changes in the average rating. In this case, the bars would be sorted from high to low with only the highest 10 increases and decreases over time being displayed. Available filters include a single-value list from which to choose the genre and a multiple-value dropdown list from which multiple movies could be displayed. Regrettably, due to each film's list of genres being included in one column in the dataset, I couldn't find a way to allow for multiple-selection, only single-selection.

Third visualization

Much like the second visualization, the the third one expanded upon the one before it, adding new context. Whereas the second visual, a bar chart, shows only the rating difference over time, this visual allows the user to visualize the direction of these shifts (Y-axis) and their relationship with a film's popularity. Much like how a film ranking poorly doesn't preclude it from being popular, neither does positive reception guarantee popularity.. In short, this visual answers three questions: 1) "is this film relatively popular" and 2) if so, "has it remained popular out of reverence or the opposite (based on direction and magnitude of direction in average rating over time)"? Thirdly, because films with a low number of ratings (closest to 0 on the X-axis) can have skewed averages, the viewer could more easily determine which films with high differences in average rating are likely to be true outliers if they have a sufficient sample size of ratings, which increases from left to right on the X-axis.

Analytics and Queries

Time series data was found to be useful for addressing, via the first visual, the question “did opinions of this film improve over time?”. It clearly displays the starting and ending points, as well as comparing the change of ratings for a specific film or films that a user would be interested in. This was achieved by creating a table function using aggregated (average rating) time series data. It required more time to figure out than expected, and was achieved by taking the absolute value of a lookup function that took the average rating in the first year of a film’s appearance and subtracted the average rating in its’ last recorded year.

As mentioned earlier, the second visual would address the question that prompted the project, mainly “which films experienced the greatest shift in ratings/public opinion?”. It was intended to serve not as a query for users to analyze one or two films in a granular fashion, but instead to quickly view outliers/extreme values and compare them either in terms of which films’ ratings rose or fell by the greatest amount within each genre, and to see whether the viewers film(s) of interest experienced as large a shift in ratings over time. In addition to rankings, bar charts still allowed for the labeling of the amount of the change over time, so viewers could still directly compare films that performed similarly enough that their bars appeared to be the same length.

The third visualization relates number of ratings (as a way to measure popularity) using the X-axis with changes in rating over time on the Y-axis. In short, this visual is used to answer three questions: 1) “is this film relatively popular” and 2) if so, “has it remained popular out of reverence or the opposite (based on Y-axis direction)”? Thirdly, because films with a low number of ratings (closest to 0 on the X-axis) can easily have skewed averages, the viewer should be able to determine which films with high long-term changes in their ratings are likely to be true outliers if they have a sufficient sample size of ratings, which increases as the viewer moves from left to right on the X-axis.

Rationale for dashboard design

When designing the first line chart visualization in particular, I aimed to provide as much information with as little ink as possible. Therefore, overall-long term changes in ratings would be reflected by the color of each line - green for net gain and red for net loss. The exact rating change over time was added to the tooltip so the viewer wouldn’t need to interpret the amount of the change using the Y-axis. Filtering would be permitted to a minimal degree, so only individual movie titles could be selected. Filtering by genre would have it’s own uses, but given the number of movies within the dataset, cluttering the line graph with the dozens of films from each genre wasn’t practical. Instead, genre filters were applied only for the second and third visualization, which already supported comparison of films en-masse.

The use of a diverging red-green color palette was carried over to the other two visuals as well, as they had at least one metric in common (difference over time) with the first visual and both the bar chart and scatter plot featured quantitative values centered around zero.

Citations

- GroupLens, & Kim, E. (2018, August 15). *MovieLens 20M dataset*. Kaggle. https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset