DSA2101 Assignment 03

Lee Pak Shuang

2022-11-11

Introduction to the dataset

The data set is provided by The Economist, sourced from IMDb. In the data, each row represents a season from a TV show (e.g. Season 3 of Better Call Saul) and contains the following columns:

Column	Type	Description	Example	Minimum	Maximum
`titleId`	character	Unique identifier for the particular season of the particular show	tt3032476	NA	NA
`seasonNumber`	integer	Season number	3	1	44
`title`	character	Show title	Better Call Saul	NA	NA
`date`	date	Date aired	2017-05-13	1990-01-03	2018-10-10
`av_rating`	numeric	Average user rating across all episodes in the particular season of the show	8.9852	2.704	9.682
`share`	num	Percentage of user reviews received out of all user reviews for TV shows on IMDb that year	1.61	0.00	55.65
`genres`	character	Sequence of genres that the show belongs to	Crime,Drama	NA	NA

Total rows: 2266

Plot 1

The goal was to visualize the distribution of ratings over time.

The base is a bubble plot, with the ratings of each season plotted against the date the season aired and each bubble had a size proportional to its share of reviews received. This allows the viewer to see how the ratings have been distributed across time:

1. Largely clustered between 7 and 9/10
2. Slightly lower and more spread out in the 1990s and more spread out in the 2010s
3. The number of seasons released increased with time
4. Bias for seasons that have been reviewed more (more popular) to be higher rated

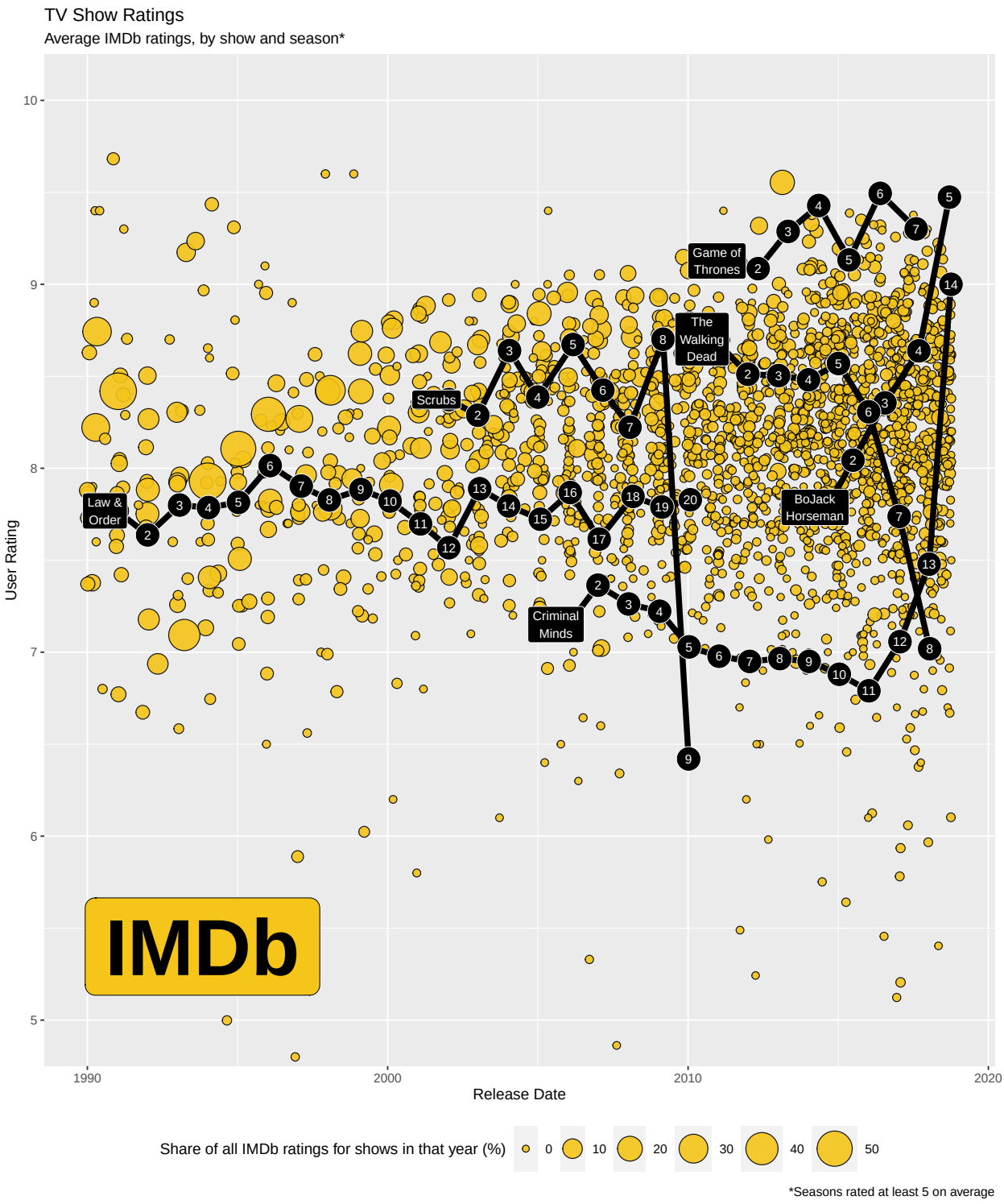
The bubbles where all filled IMDb's yellow for the aesthetic appeal, coloring the bubbles according to primary genre was explored but was not meaningful, it just made the plot ugly. A black stroke outlining each bubble, nudged positioning, and a slight transparency was used for the bubbles to enhance the clarity of bubbles and improve the visibility of overlapping bubbles. The chart was constrained between 5 and 10 on the y-axis as there were incredibly few (about 10) bubbles below 5/10, the distribution would be too squished if the entire y range was plotted.

A curated selection of shows were highlighted, these shows were chosen as they reflected a variety of "paths" that shows have taken in terms of reviews: from consistent quality like "Law & Order" to wild moves like "Scrubs". This was achieved through a combination of line, point, text, and label geoms. Black was chosen as it provided good contrast and was consistent with the IMDb theme. The IMDb logo was emulated using a label geom for decoration.

This plot is intended for a layman audience. The design emphasis is for visually delivering an attractive visual that quickly gives the viewer and overall impression of the data. The plot is low in dimensionality and density while high in figuration and decoration.

Variables:

1. Ratings of seasons
2. Release date
3. Share of ratings received
4. Change in rating across time and seasons (curated)


TV Show Ratings
Average IMDb ratings, by show and season*
User Rating
Release Date
Law & Order
Scrubs
Criminal Minds
Game of Thrones
The Walking Dead
BoJack Horseman
IMDb
Share of all IMDb ratings for shows in that year (%)
0
10
20
30
40
50
*Seasons rated at least 5 on average

Plot 2

The goal was to expand on the idea of the rating "paths" that shows in Plot 1, demonstrating what the nature of these "paths" are and how they vary depending on the ratings of the pilot season.

The base is a line plot, with the ratings of each season plotted against the season number (not time). The last season of each show is plotted with a diamond. However, plotting this on a single graph would be very crowded. The plot is faceted according to binned levels of pilot season ratings so that the viewer can examine each bin individually. This allows the viewer to see how the ratings of shows evolve across seasons:

1. The few shows that start at 9-10 mostly do not get even better
2. Shows that started at 8.5-9 mostly maintain their quality, but declined after about 5 seasons
3. Shows that started at 8-8.5 mostly continued to stay within 7.5-9
4. Shows that started at 7.5-8 mostly continued to stay within 7.5-8 or improved to 8-9
5. Shows that started at 7-7.5 mostly improved to 7.5-8.5
6. Shows that started at 6-7 mostly improved to 7-8
7. Shows that started at 5-6 mostly declined to 4.5-5.5

The plotted diamonds that indicate the last season of each show display the distribution of total seasons as well as serve as a visual marker of the end of each "path" for interpretability. The median of these diamonds is drawn using a vertical line that is behind the colored "paths" to prevent any obfuscation, it serves to show the trend in number of seasons produced depending on pilot season ratings. this is important because due to the sheer density of single-seasoned shows, the medians are much lower than one might anticipate from the colored "paths" alone.

In order to facet the plot without losing context, the scales are kept the same and the "paths" from other facets are plotted in grey as well so the viewer can see where the colored "paths" exist in relation to the entire set of "paths" in grey.

To make it easier to resolve the individual "paths", the color of the "paths" is mapped to the total number of seasons eventually aired by the show. The mapping is from IMDb yellow to generic purple as the total number of seasons aired increases, this makes the longer lines more outstanding.

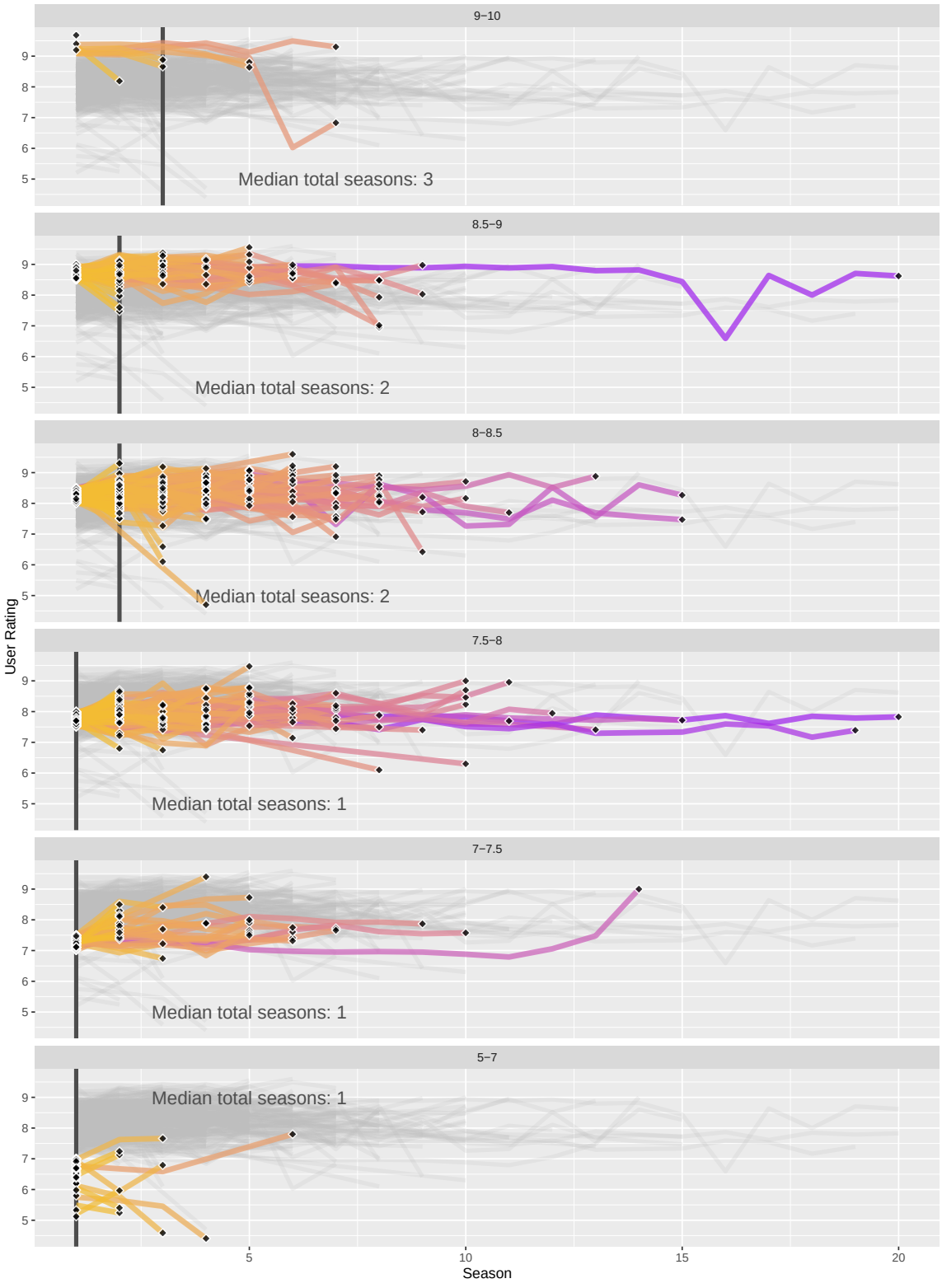
This plot is intended for a layman audience. The design emphasis is to build upon the viewers curiosity from Plot 1 and allow them to explore the richness of the "paths" through Plot 2, encouraging exploration and the development of their own conclusions about the data. The plot is low in dimensionality and while high in figuration and redundancy.

Variables

1. Ratings of shows by season
2. Season
3. Total seasons aired

Are TV pilot seasons' ratings any indicator of subsequent seasons' ratings?

Average IMDb ratings across seasons, by show and rating of pilot season



*Note: Some shows have not stopped producing new seasons