DSA2101 Assignment3

Veronica Angelin Setiyo (A0240487B)

2022-11-10

Intro to dataset: Dr. Who

The dataset I had chosen is Dr.Who dataset. This dataset has 4 sub datasets, all about the episodes of Dr. Who series. directors.csv and writers.csv have a list of directors and writers respectively, with their corresponding story_number(s) that they worked on. episodes.csv has data for episodes in season 1-13(partial), which includes rating, story_number, season_number, episode_number, and other data. imdb.csv has very similar information to episodes.csv. In my exploration, I decided to use imdb.csv's ratings instead of episodes.csv since imdb.csv also has data on the number of ratings received for each episode, which I wanted to explore as well.

Questions and Hypothesis

1. What is the popularity like for the different Dr. Who seasons?

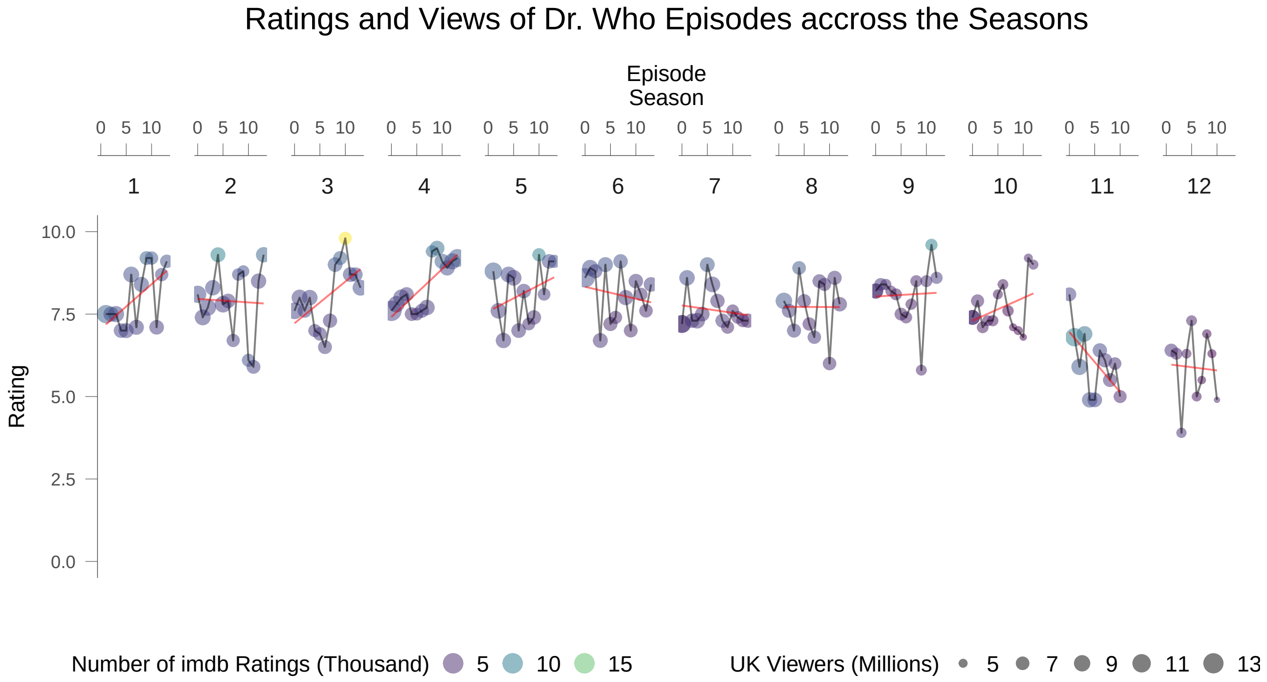
My hypothesis is that there are some seasons that are more popular than others in terms of ratings or views. Listening to some news article, the later Dr. Who seasons are not as good as the earlier ones.

2. Can we differentiate between good writers and directors?

My hypothesis is that some writers and directors are more successful in making episodes that sell. Some of them might consistently do well while others might consistently do badly. Good writers combined with good directors should also produce good episodes, vice versa for bad ones.

Note: I had a lot of trouble transferring the plot from R Script to knitting it at Rmarkdown without the plots jumbling up horribly, which is why I decided to save the produced plots as images, then also including the saved images in this pdf document. The code I used to produce the graph in R Script is inside the R Markdown document.

Plot 1



Insights:

There is some variability both in ratings and views fir the episodes within each season, while number of ratings received is more consistent with rare anomalies such as that in Season 3.Some seasons have a general trend of increase in rating while others have a decrease in rating.The greatest decrease in rating is in season 11 while the greatest increases are in season 1, 3 and 4.The last 3 seasons did poorly, with lower ratings and viewerships, except for season 11 which still had considerable viewership despite the low ratings. Season 3 episode 10 has the best rating, with the highest number of ratings received. After googling, it is said to be the best episode of Dr Who.

"'Blink'(Seson 3 episode 10) is one of those episodes – one of the best of the new series, and a fantastic 'stand-alone' that pretty much sums up both the intricacies of the time lord world and the abilities of the Doctor." -an imdb review

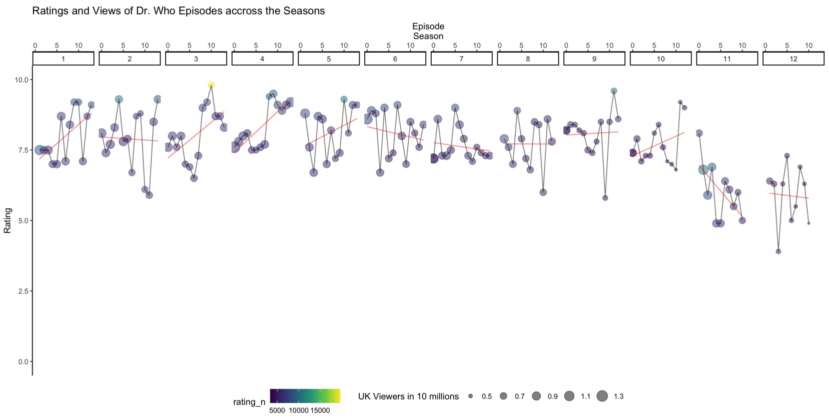


Figure 1: Plot 1, image saved from R Script version, which have not undergone further edits

Intended audience:

Since the data is not updated and the scales are not precise, this plot is best suited to be seen by laypeople who are interested in knowing the general trend of the performance of Dr. Who episodes throughout the seasons.

Design choices:

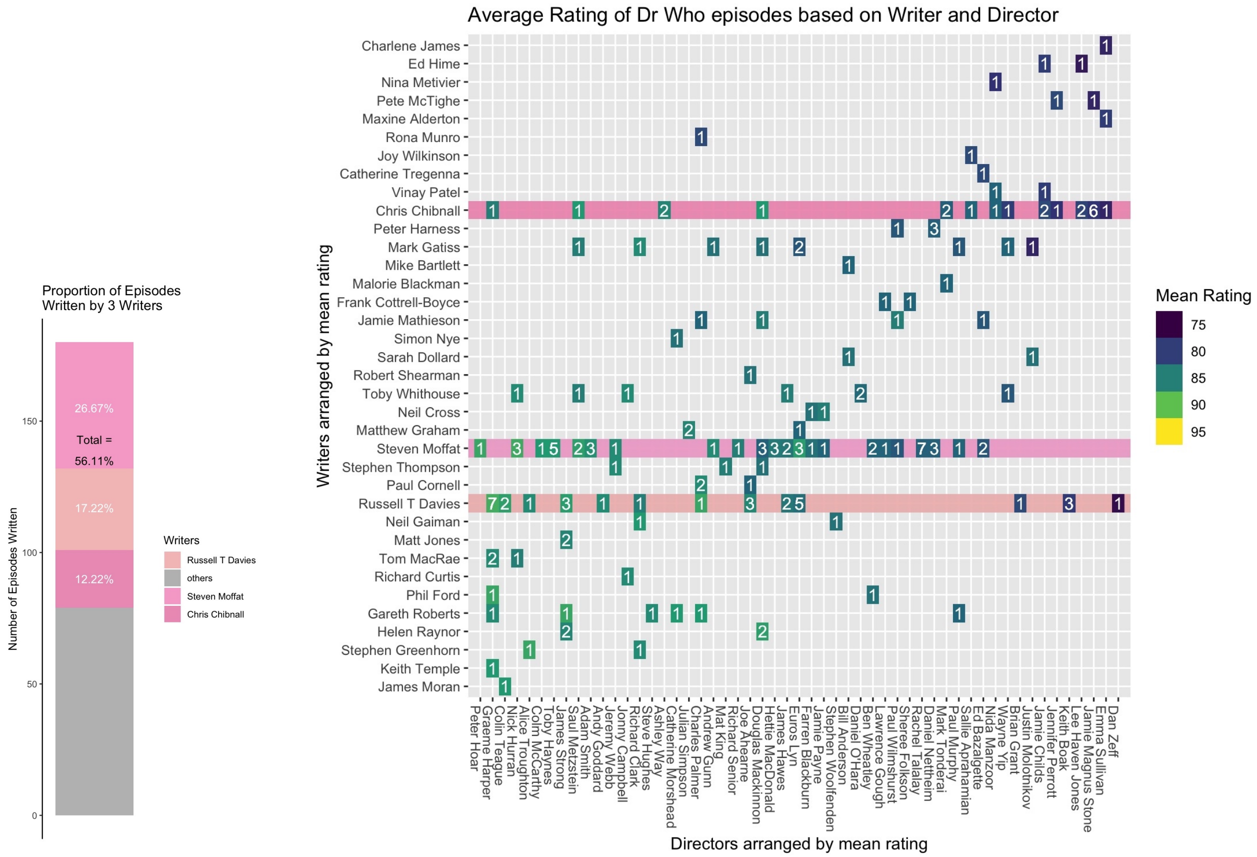
The geom_points are semi-transparent to allow the viewers to see overplotting, which we can also see in episode zeros of seasons 7, 9, and 10 as they have multiple episode zeros. I used viridis scale colour to create good contrast between the number of ratings, which successfully highlighted the number of ratings received in Season 3 Episode 10. I also used faceting all in 1 row to allow the users to compare the ratings across the seasons, which would not be possible if I had faceted them in rows and columns. Following Tufte's advice, I minimalised all the border lines, reducing their thickness to avoid cluttering and distraction from the data, and removed the background grid. I resized the texts and other elements so that the data can fit snugly despite having so many points.

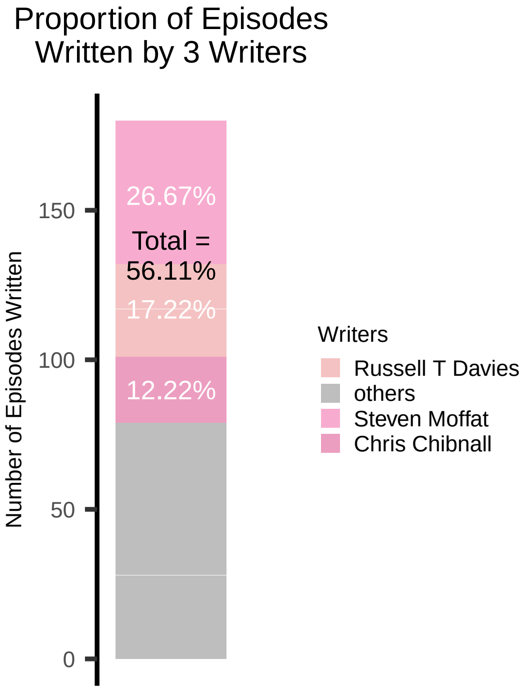
What I would improve on had I had more time:

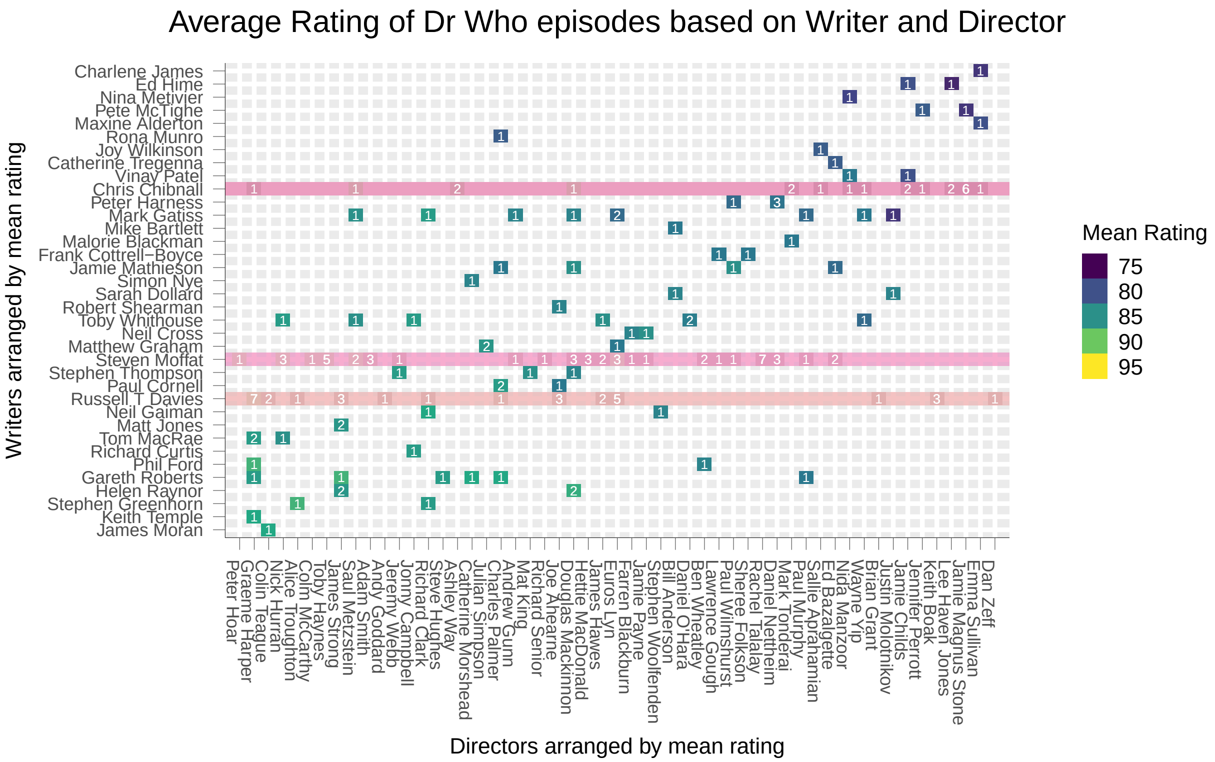
The legend should be in continuous format such as that in the embedded image which showed the whole range of possible colours, instead of what is produced from the R Markdown knitting. I could not figure out how to convert it back despite spending days on it. I would also vary colour the linear model lines: red for decreasing trend and green for increasing trend.

Plot 2

This plot includes 2 sub plots: the bar graph and the raster graph. The bar graph is a continuation of the raster graph. The embedded image (fom R Scipt) successfully shows the fill plots, which is why I hope I will be graded on the embedded image since it seems like the resolution limitation for graphs in R Markdown is what caused the fill plots to be unsuccussfully shown. Based on this site, it might be due to my pdf viewer.




Proportion of Episodes
Written by 3 Writers
Number of Episodes Written
150
100
50
0
26.67%
Total =
56.11%
17.22%
12.22%
Writers
Russell T Davies
others
Steven Moffat
Chris Chibnall



Insights:

Focusing first on the raster graph, as the writers and directors are ordered by the mean atings of the episodes they worked on, we can see that the closer they ae to the bottom left corner, the better their performance was. From here we can also see some of the directors who dragged down good writers. For instance, writer Russell T Davies has quite good portfolio with quite good mean ratings working with several directors. However, the episode he did with director dan Zeff did quite badly relative to other episodes.

Some of the most popular writer-director combinations are Russell T Davies-Graeme Harper and Steven Moffat-Rachel Talalay, each with 7 episodes. From the mean rating, we can also see that Russell T Davies-Graeme Harper episodes do better than Steven Moffat-Rachel Talalay episodes.

Also, looking at the numbers alone, I could see that some writers did a lot of stories. I managed to confirm this by making the col plot. The 3 writers, Russell T Davies, Steven Moffat, and Chris Chibnall alone contributed to writing 56.11% of the stories in Dr. Who up to season 12.

Intended audience:

Although the rating scale is not very precise, by ordering the writers and directors based on their mean performance (metric is mean rating of the episodes they worked on), producers can then compare and contrast to see which writers are good and which directors are good, and what are some of the best combinations. For laypeople, I would say the graphs are understandable but would not be particularly useful.

Design choices:

The raster graphic is indeed quite sparse (as I have consulted with Prof Vik). However, after faceting based on season, I could no longer find the overall insights as I could with the full raster graph despite it being rather sparse. Thus, I kept the sparse raster graph format. Each plot is accompanied by the number of instances at that point to handle the over-plotting.

I used viridis scale colour to create good contrast between the mean ratings. In highlighting the 3 rows of writers who worked on the most Dr. Who episodes, I chose pink shades that has good contrast with the viridis colours yet is not overpowering. The pink shades are similar enough to create a holistic clump of the 3 writers in the column graph, creating good contrast with the "other" writers which I did not want to highlight, hence I used a muted shade of gray. I also kept the background grid in the raster graph to aid the users to trace each plot to either its corresponding writer or director. The x axis labels of the raster graph are tilted to avoid clustering, producing a cleaner graph, while the x axis of the column graph was deleted as it served absolutely no purpose.

What I would improve on had I had more time:

The legend on the right should have a continuous colour gradient. It was a continuous colour gradient when I first made it, however after multiple edits it became discontinuous and I did not have enough time to figure out how to convert it back. I would also reorder the column graph and its legend keys to match the order in the raster graph so that it's more intuitive for the reader to see the link. The sizing could be improved as well, and I would also love to highlight the 3 writers' names in their respective pink shades.