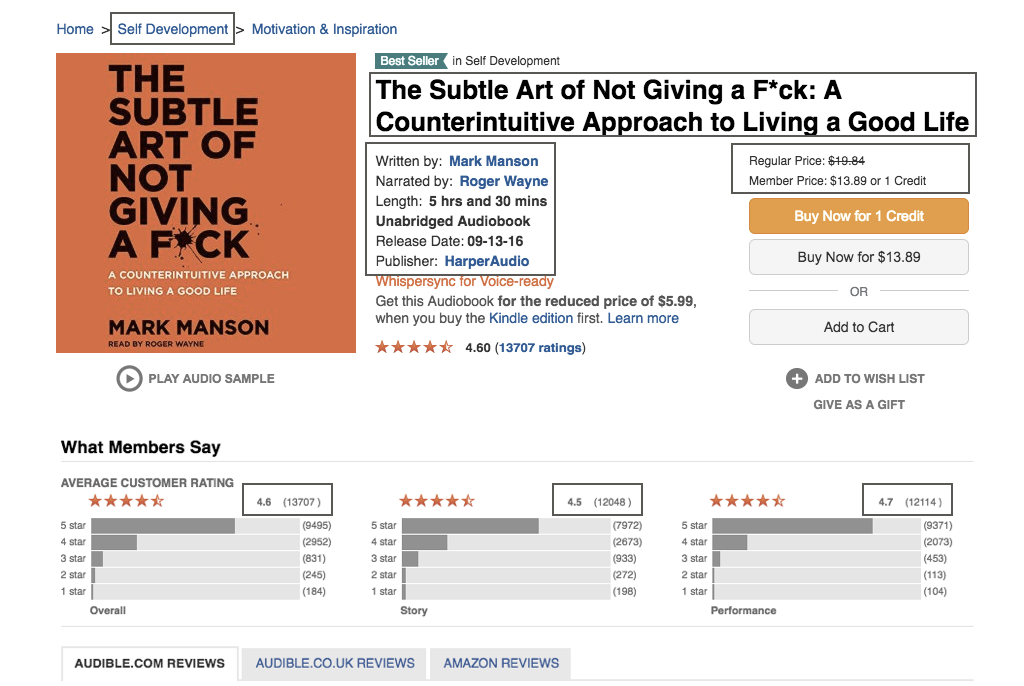
Drama vs. Words: Does the story or the performance determine the rating of an audiobook



In 2016, the [Wall Street Journal](https://www.wsj.com/articles/the-fastest-growing-format-in-publishing-audiobooks-1469139910) wrote that audiobooks were the fastest growing format in publishing. As an effort to understanding this segment of the publishing industry I analyzed English-language audiobooks to determine if the overall rating for an audiobook is determined more by the performance or by the story. The data is from the Audible.com website.

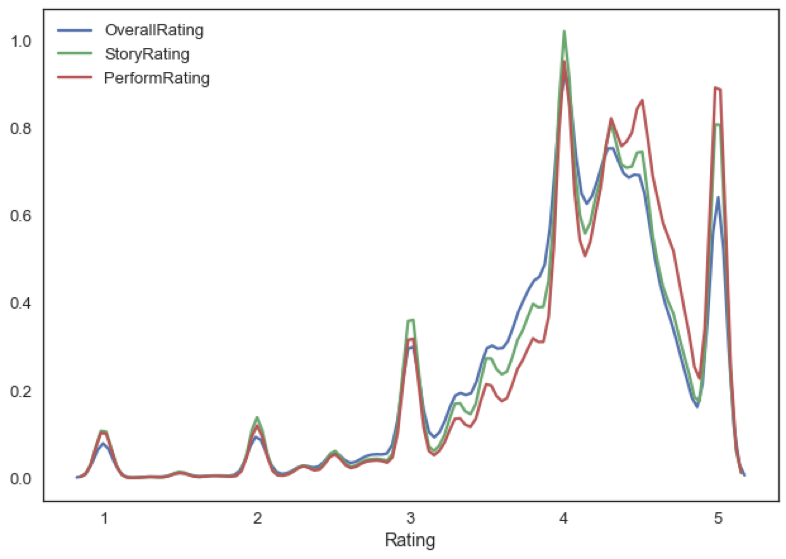
Audible.com is a technology company that produces, sells, and markets audiobooks. The audiobooks from Audible.com are purchased through its online platform and can be streamed or downloaded through its proprietary app. Audible was established in 1995 and has over 150 employees. Audible was purchased for $300 million dollars by Amazon.com in 2008. Audible has a library of over 200,000 English-Language audiobooks. This relatively long history and deep catalogue makes Audible a good source of data for audiobooks. The webpages for audiobooks also have a uniform format easing the difficulty of scraping.

Using the Scrapy framework, I scraped data from the Audible website on English-language audiobooks. This table was used to populate a table that included data about the book along with the ratings and number of ratings. The data collected from the page is shown below.

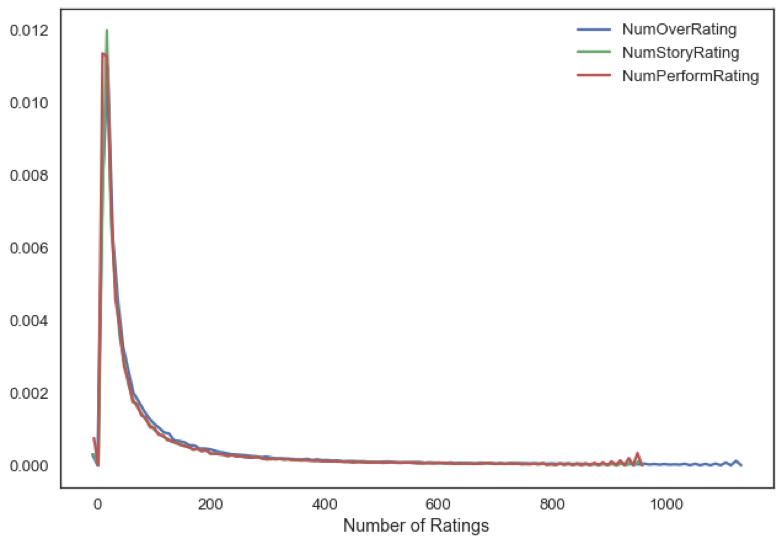


The scraper was setup to start on the first page listing links audiobooks. Each audiobook link was followed and the data downloaded and returned to the original page. The link to the next page was followed using recursion until the final page is reached. Using this process, a table of ~203,000 audiobooks was generated. Of these ~42,000 audiobooks were never reviewed. With several hundred more having only Overall reviews. Only books with complete reviews are included in charts with reviews, while the entire dataset is included for charts characterizing the data.

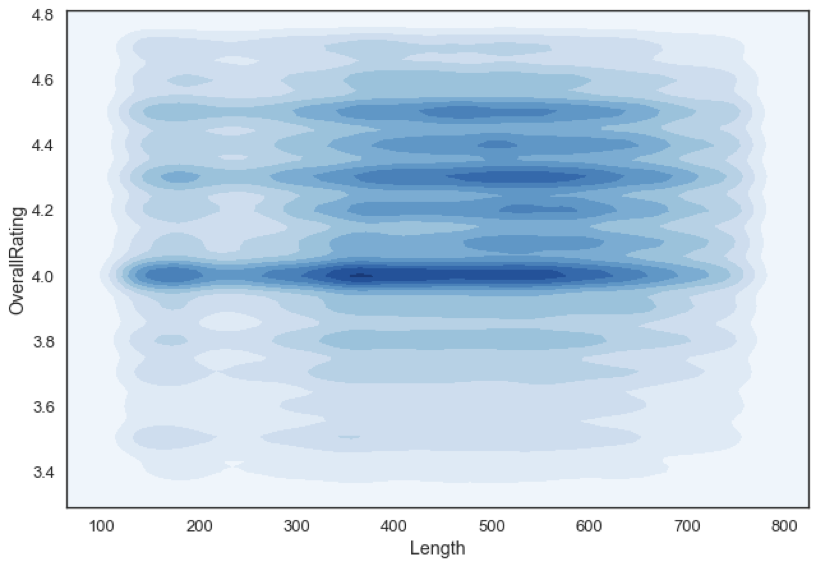
The distribution of the overall rating, performance ratings, and story ratings are shown below. The most common rating for an audiobook is 4 for all categories. The second most common is 5 for the performance rating and between 4 and 5 for the overall. These indicate the ratings are not equally distributed between the lower ratings are not given often. Thus, most rated books should be skewed high. Since people do not usually purchase books they do not expect to enjoy this is not surprising.



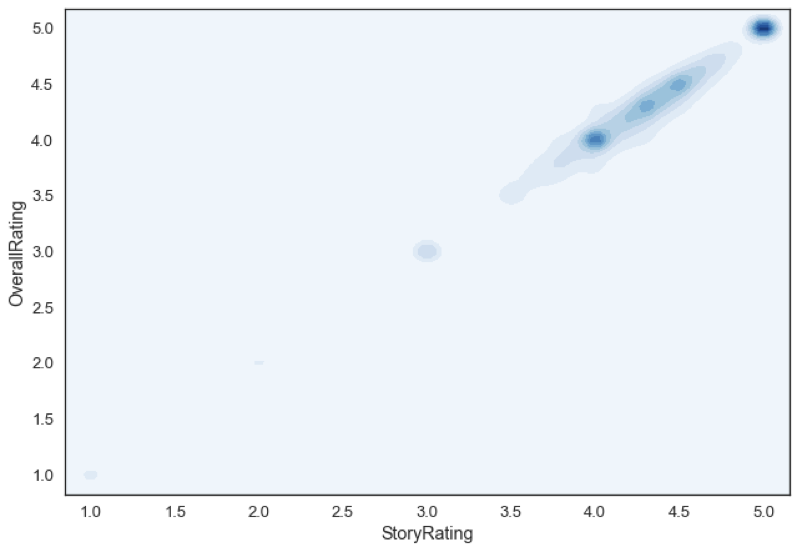
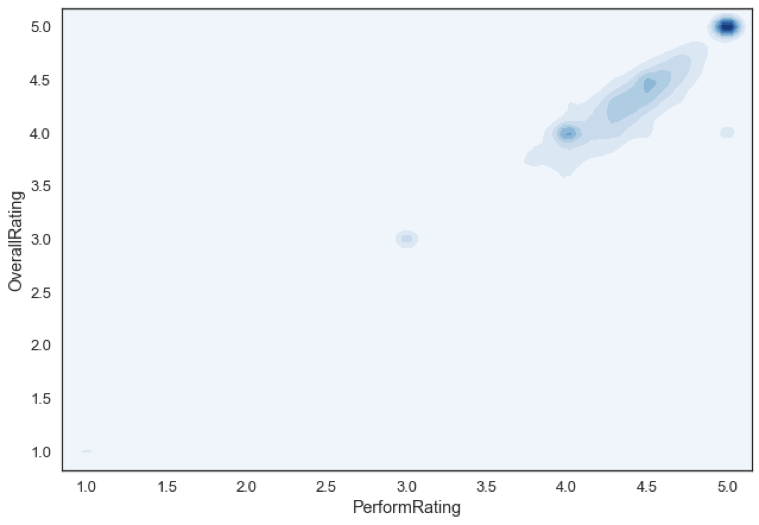
The number of ratings per book shows a sharp peak at 1. Of the ~150,000 rated books on Audible, most have a single rating. The low number ratings per book indicate why the distribution of ratings (above) has strong peaks at the integer values. Note that the rating distribution (below) removes all ratings over 1500 for clarity.



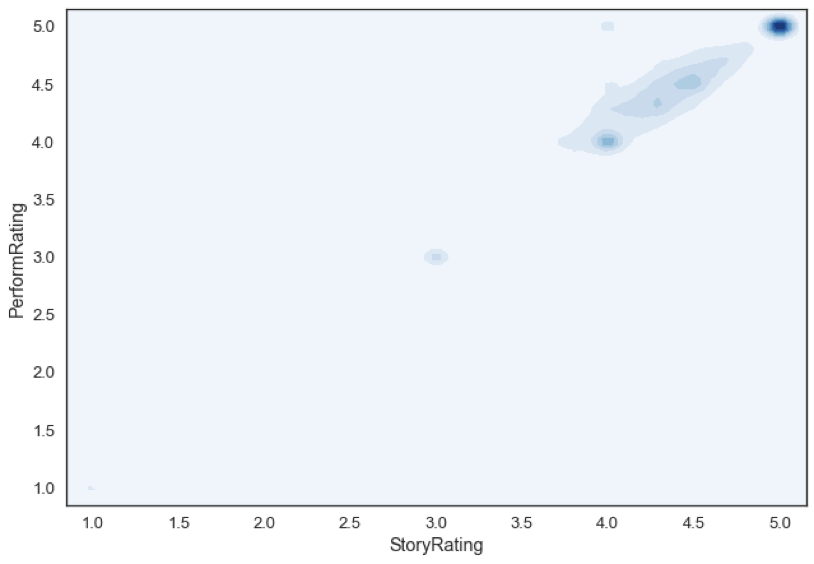
Book length is a possible important measure, longer books may be rated lower due to length. I plotted the density of the overall rating vs. the length in minutes (below) with the outliers removed (2 standard deviations from the mean length). From the plot, it is easy to see that longer books are not rated lower than shorter books. The trend is reverse with longer books having a higher rating than shorter books.



We can plot the overall rating against the performance rating and the story rating using a density plot to examine the relationship between these variables. As seen below, there is a very strong linear relationship between these variables. This correlation indicates both the story and they performance are correlated with the overall rating. This can be confirmed by the correlation coefficients, overall-performance 0.81, and overall-story 0.88. Once could tentatively conclude that while both are important, the story is more correlated with the overall score and is more important.



To future explore the validity of the result I checked the correlation between the performance and the story ratings. This value also shows a string linear relationship and has a high correlation coefficient of 0.78. This correlation indicates that one cannot make a definite determination based on the correlation to either performance or story as one variable could determine the other.



To understand which is more important we need to consider cases where one is held constant while the other can vary. One method is to find books with multiple narrators. The stories in these books don’t vary but the narrators are different. I explored 158 books with multiple narrators. These books have overall, story and performance rating that are almost same for each book but vary with narrator. It only a few cases did the ratings diverge. This we can conclude that the perception of the story is determined by the performance and the performance rating helps to determine both the overall and the story rating. Writes should carefully consider who narrators their story to have a rating the story deserves.