Ilyas Jaghoori

Professor Evan Bagley

APMA 3100

May 2nd, 2023

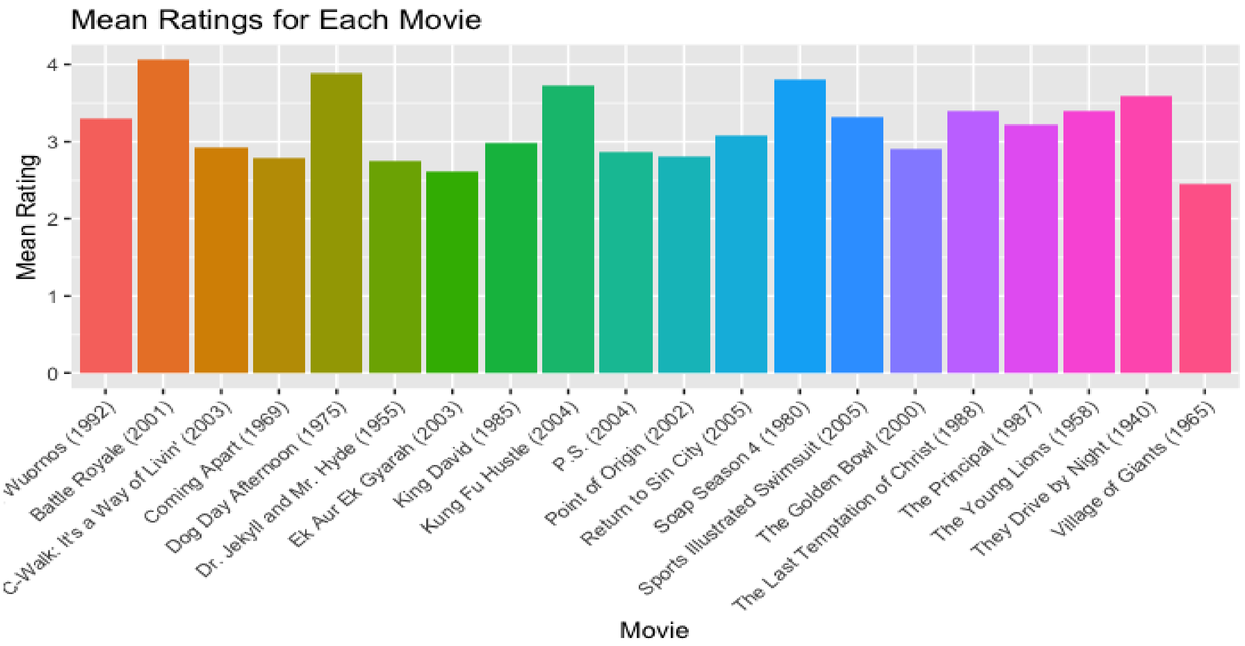
Introduction to Data Science Final Report: Movie Ratings and the Impacts of Seasonal Changes

On my honor as a student, I have neither given nor accepted any unauthorized aid on this assignment or exam.

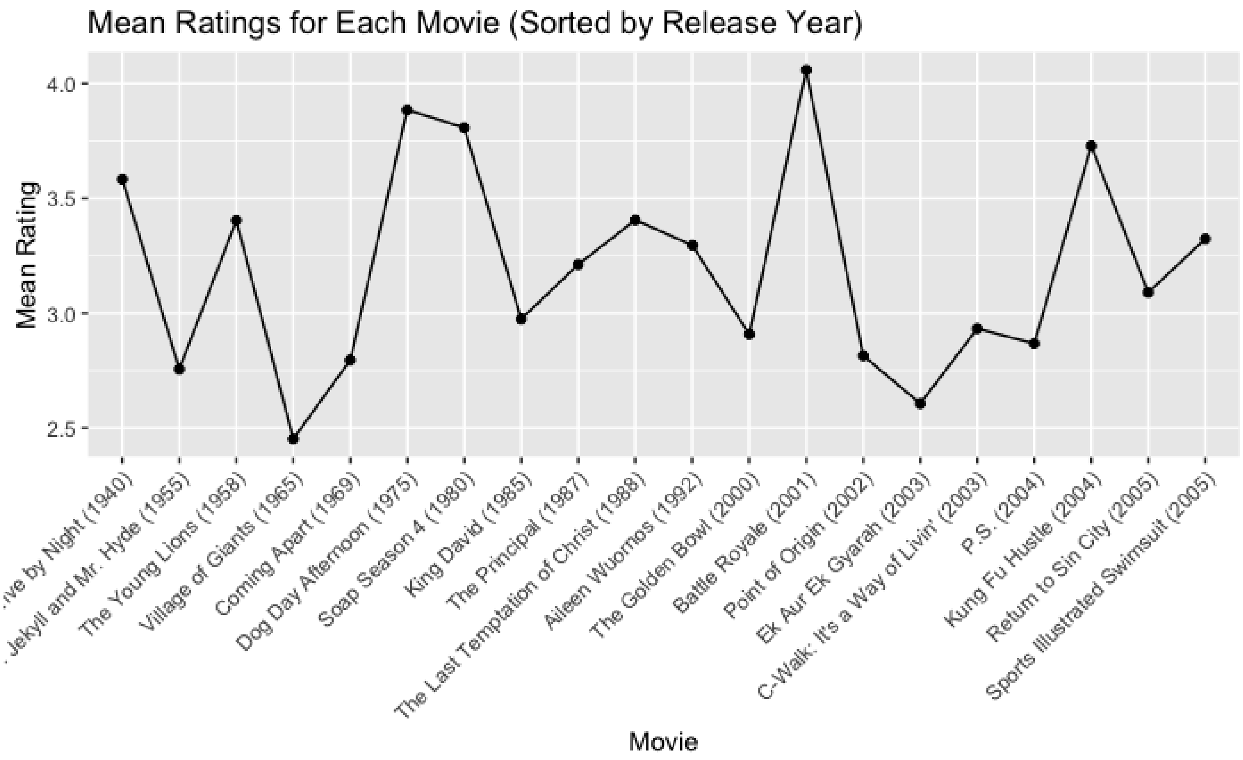
-Ilyas Jaghoori

Movies throughout time have always had critics, regardless of how acclaimed they may be. An analysis of ratings for 20 randomly chosen films spanning various periods of the modern era was conducted to identify potential trends in diverse aspects of their information over time. My goal was to achieve a better understanding on what different factors could potentially change the ratings for these movies or the amount of ratings received per movie. The selected movies encompass a wide variety of different genres such as action, documentary, romantic comedies and even Dramas. This vast range of genres really provided me with a more accurate understanding of overall trends in cinema. By examining movies from different genres and time periods, I aimed to gain insights into how audience perceptions and critical evaluations may vary across diverse film contexts. After an exhaustive data visualization process and different correlation analyses, I was able to notice that seasonal changes impacted the number of movie ratings critics gave and the overall mean ratings.

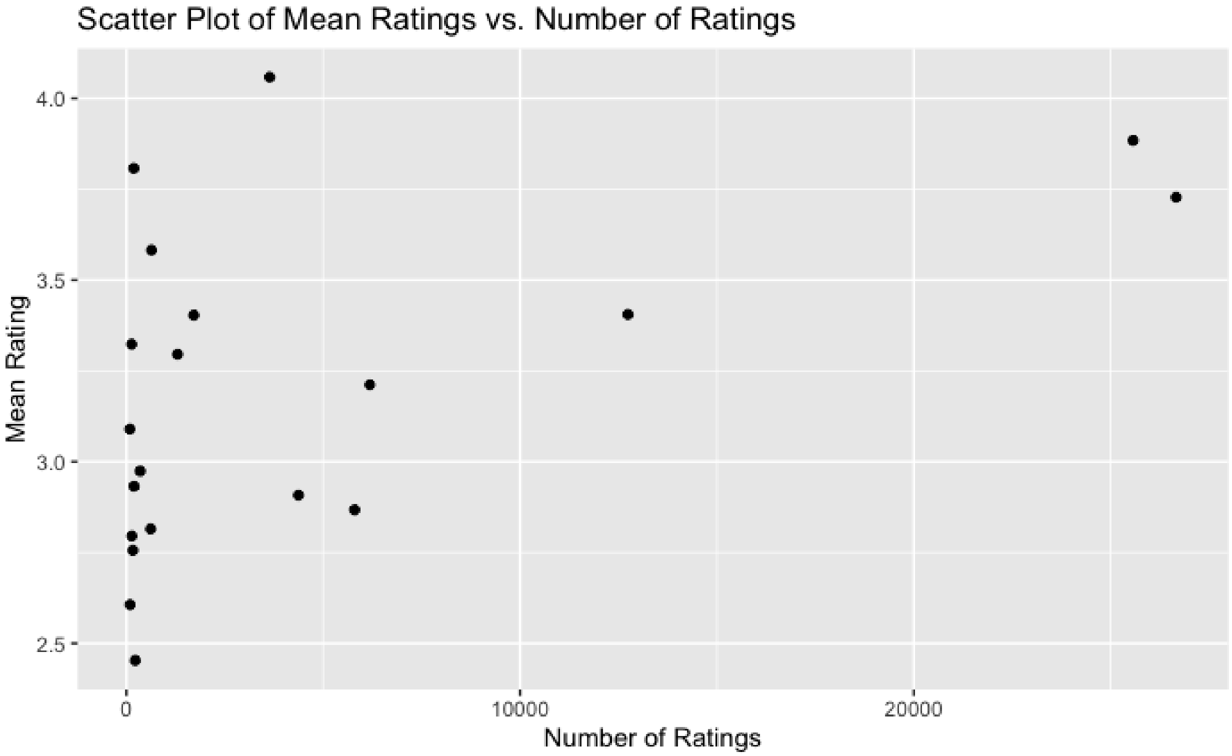
To begin the experiment, I started off with understanding what each movie had as their mean rating. This can be visualized with a bar-plot shown below:



As seen, the movie with the highest mean rating was Battle Royale (2001) with a mean rating of 4.058. In last place, the movie with the worst mean rating was Village of Giants (1965) with a mean rating of 2.453. I concluded this initial experimental analysis believing that mean ratings started off very low in the earlier years of the movie sample but slowly rose throughout time. This, however, was disproved. I created a line graph that had the movies sorted chronologically on the x-axis and the associated mean ratings for the movies on the y-axis. The graph can be seen below:

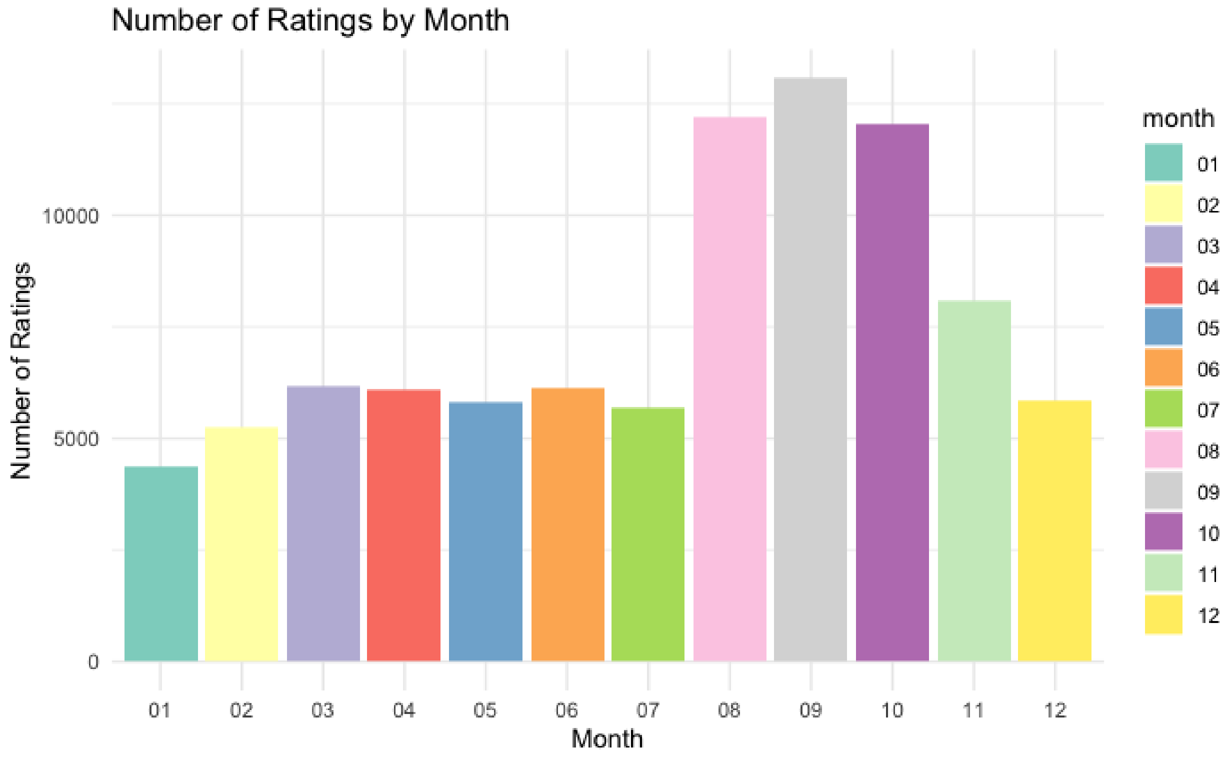


Shown here we see a very evident fluctuation in the movie ratings over time. No one period of time follows a trend-like pattern, either positive or negative. This graph disproved my hypothesis that mean ratings increased over time but did help me visualize the diverse nature of movie ratings across different time periods. After concluding that initial hypothesis was false, I moved on to analyze another potential trend, how correlated the mean ratings of the movies were with the amount of ratings that each movie received. Mentioned earlier, the dataset given in the beginning of the project included 20 movies; however, each movie had a different amount of ratings. For instance, Dog Day Afternoon (1975) received precisely 25,568 ratings, coming in just behind Kung Fu Hustle (2004) who had 26,662 total ratings. On the other hand, the lowest number of ratings was Return to Sin City (2005), with a mere 89 votes. This shows the significant variation and potential skewness in the information available about these movies. A correlation analysis was conducted to see the potential impact the number of ratings had on the overall mean of the movie ratings. After running the code and creating a scatter plot that can be seen below, it was found that the two variables had a 0.5058236 correlation coefficient. This demonstrates that although not strong, there exists a correlation between the two variables such that as the number of ratings increase, so do the mean ratings of that movie.

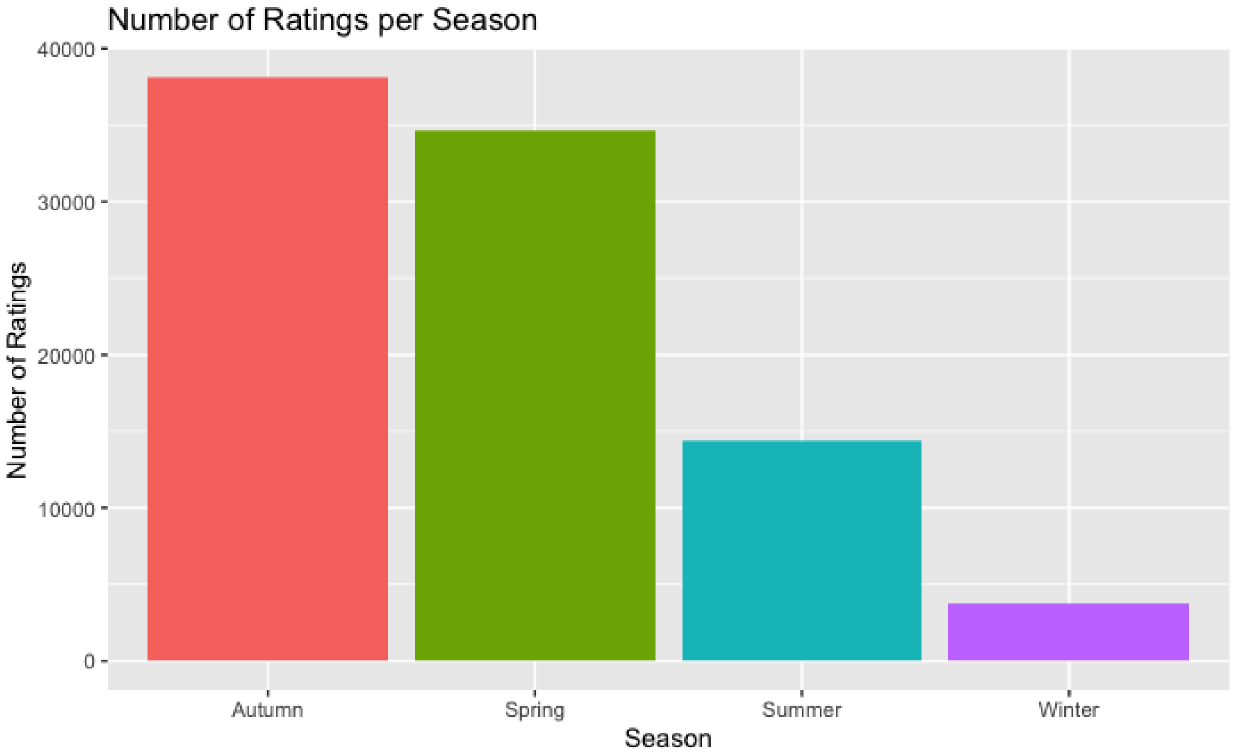


The main takeaway from this specific analysis is that extreme differences in the number of ratings can cause some issues when looking at data. Movies with more ratings might give us more trustworthy and representative feedback, while movies with fewer ratings could skew the data because there aren't many ratings to go by. It is important to be careful when drawing conclusions from this kind of data.

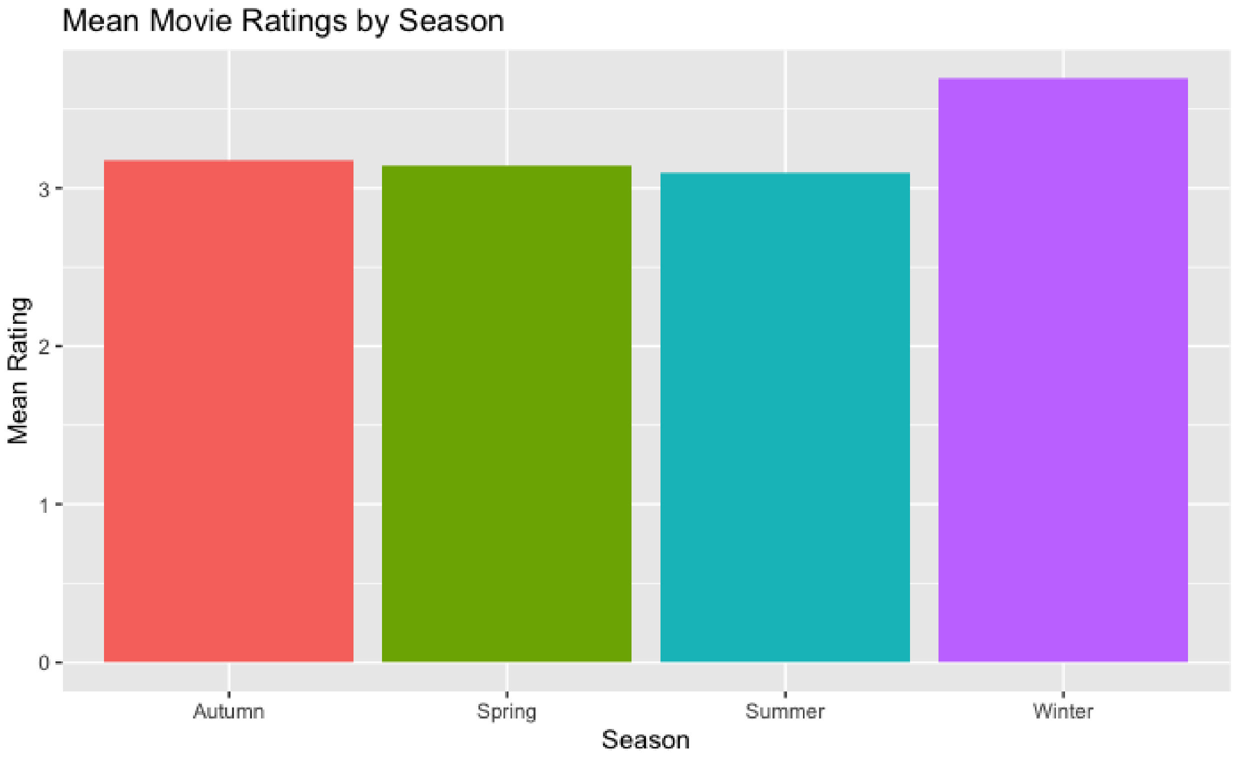
After conducting an extensive analysis of factors like movie run time, genres, and release dates, it was ultimately determined that distinct trends were present in seasonal changes with regard to the number of movie ratings presented. This trend can be attributed to multiple factors, such as movie release dates during specific seasons, critical preferences, and audiences preference to watch specific genres only in particular seasons. Below, a visualization of the number of ratings per month can be found:



It can be seen that there is a significant leap from the month of July to August and then a miniature plateau at that high level of ratings through the month of October before the ratings drop to their normal levels in November and December. This shows that with the data given, the critics were more likely to give ratings during the months of August-October. This surge in ratings could be attributed to different factors such as anticipated releases, movie festivals or the prevalence of certain genres during these specific months. These months don't technically have similar seasonality as the month of August is more associated with Summer and the months of September and October are more generally considered to be Autumn months. Due to this, an analysis was conducted to see if associated seasonal months have any impact on the data. In this experiment, the months of spring were considered to be March, April, and May; the months of summer were considered to be June, July, and August; the months of Autumn were the months of September, October, and November; and the remaining months were considered Winter. Below, a bar-plot graphical representation of number of ratings per season can be found:



When comparing this plot to the previous bar-plot, the data holds to be consistent; the months of Autumn were the months where there existed the most ratings. What I found really interesting about this part of the visualization was how much of an impact a low sample can have on the analysis of different variables. For example, it can be seen that the season with the least amount of ratings was winter with a count of 3,772 ratings. Now, when doing an analysis of the mean rating per season, it was no surprise that winter led the way with the best mean rating. This can be seen below:



The low amount of ratings creates an environment where numbers that are more skewed in a positive direction have a much stronger impact on the overall mean rating compared to those with a larger sample. It would be a lot harder for a high rating to have a significant impact on the Autumn mean rating as Autumn had about 38,096 ratings. This shows that the conclusion made earlier about approaching findings with caution is always important when analyzing data. It can confidently be stated that movies that were released in the Autumn have significantly more data to analyze and also a better mean rating compared to the other seasons.

To conclude, it was observed that trends in movie ratings can be more complex than initially assumed. It was important to dive deeper into the data instead of analyzing the surface level and to also take caution when approaching conclusions. Stakeholders who are interested in finding out about changes in movie trends can find this information useful when understanding that seasons may have potential impacts on the way critics perceive the movie.