Dong Woo (Michael) Kang

Steven Green

DU-VIRT-DATA-PT-12

Project 1 Written Analysis

**Research Question**

The question for research that I inquired about was the relationship between the User IMDb rating and Critic-centered Meta-score for the IMDb Top 1000 movies. How do the two compare and if a relationship exists, what kind of relationship is it?

**Introduction**

First a little background, the IMDb is an online user-inputted internet movie database. In fact, the acronym IMDb stands for Internet Movie Database. It contains information related to everything film and cinema; and the information can be inputted by online users with very little restriction. The IMDb Top 1000 movies is a unique and specific dataset that contains the IMDb 1000 top-rated movies from over the years and information relating to each movie’s public receptance, including the two variables at study: User IMDb rating and Critic-Centered Meta-score. User IMDb rating is a score that is summarized out of a rating scale from 1-10 that registered IMDb users can rate a movie title/series out of. These scores are aggregated and then averaged out to represent the total IMDb rating of a select movie title/series. The Critic-centered Meta-score is a bit more difficult to understand; this is a numerical score that is calculated by a third-party company named Metacritic and is used by IMDb to show a more focused and concise rating by professional movie critics. Metacritic calculates Meta-Score by aggregating and averaging out a percentage score inputted by a sample population of incredibly respected movie and film critics. In order to be a part of this sample population of highly respected critics, Metacritic enacts a rigid code of expectation that critics must meet before being accepted into the sample.

**Data-Cleaning and Organizing**

Before conducting any cleaning, I imported my necessary Pandas libraries and commands into my Jupyter Notebook script. I then read in my raw data set that we obtained from Kaggle that was titled “imdb\_top\_1000.csv” with the necessary Pandas library commands. Following that I immediately removed columns: “Poster\_link” and “Overview” as we agreed as a group that we did not need data containing movie summaries and online links to IMDb websites. After removing those selected columns, I dropped any rows of data that contained null values for any of the remaining columns and that brought our total rows of data to 714 total data inputs. Now that we had a cleaned set, I went on to rename columns to be more legible and coherent. It was also discussed as a group that numerical analyses will be taking place so I converted all numerical columns to float types so that analyses could be correctly conducted in Pandas, I further added Pandas formatting to these numerical columns to more clearly depict what each column represents. I copied this finalized data frame to a new CSV file and began conducting my analysis.

**Data-Analysis: Results of User IMDb Rating VS Critic-centered Meta-score**

I began my research by comparing the average User IMDb Rating to the average Meta-Score of the entire dataset of IMDb top 1000 movies. I then wanted to know which values were given out the most for the dataset of IMDb top 1000 movies and so I retrieved the mode for both columns. What I found was that the average User IMDb Rating (7.937) and average Meta-Score (77.1582) were both slightly higher in value than the mode for each respective column; User IMDb Rating mode (7.7) and Meta-Score mode (76.0). I also found out that IMDb Users and Metacritic selected critics had a general approval rating of between 70%-80% for the top 1000 rated movies in the IMDb database.

I then wanted to see if this correlation of datapoints meant causation, in that, User IMDb Rating has a notable and significant impact on Critic-Centered Meta-Score. I first used a scatterplot and then I used a scatterplot with a calculated regression line to visualize this relationship. It is also important to note that because we are looking at IMDb’s top 1000 movies, no IMDb score or Meta-score was found to be below the IMDb rating value of 7 and Meta-score value of 25. However, what I found was that while there is correlation between User IMDb Rating and Metacritic’s Critic-centered Meta-score, there is little to no causation, in that, the variation in User IMDb Rating has little to no influence on the variation found in Metacritic’s Critic-centered Meta-score. If you look at the scatter plot without the regression line, one can see that the majority of datapoints are skewed to the upper left of the data set signifying that a majority of movies scored on the lower end by IMDb Users were actually scored relatively higher by Metacritic Critics. In addition, when analyzing the scatterplot with regression calculations, one can see that while a positive relationship exists (slope value = 11.88) and the regression line depicted represents a positive correlation between the two scores, the calculated r-value (0.0789) signifies that only 7.89% of the variation in Metacritic’s Critic-centered Meta-score is caused or explained by the variation in User IMDb rating.

**Conclusion**

While it is easy to observe that User IMDB Rating and Metacritic Critic-centered Meta-score have a positive correlation, after further research, it is revealed that, in this case, correlation does not mean causation. While higher IMDB User Ratings usually do correlate with higher approval from Metacritic Critics, it is statistically proven that the Metacritic Critics do not take into account User IMDB Rating when calculating their own respective scores of a selected movie title or series. It is also good to note that because we are looking at the Top 1000 rated movies (and not some random movie data set sample) so the scores we are looking at and analyzing are on the high end of the spectrum.