IMDb Top 1000 Movies List – Exploratory Data Analysis

Topic: Everyone has their own favorite movies that hold a special place near and dear to their hearts. The IMDb Top 1000 movies is a list that many will look to for movie ideas on a regular basis. On IMDb, films are given both an “IMDb Rating” and a “Meta Score”. The former comes as a result of one-time votes submitted by IMDb users on a 1–10 score. The latter is a score given by movie critics on a 0–100 scale. This particular list ranks the movies by their IMDb rating. With about a hundred years' worth of movies to consider, I was curious as to how the thousand best movies are distributed over that time. Furthermore, I wanted to see how movie critics and regular people rated the older films versus the newer ones. Thirdly, I was curious how the movie length and gross revenue are related to the rankings.

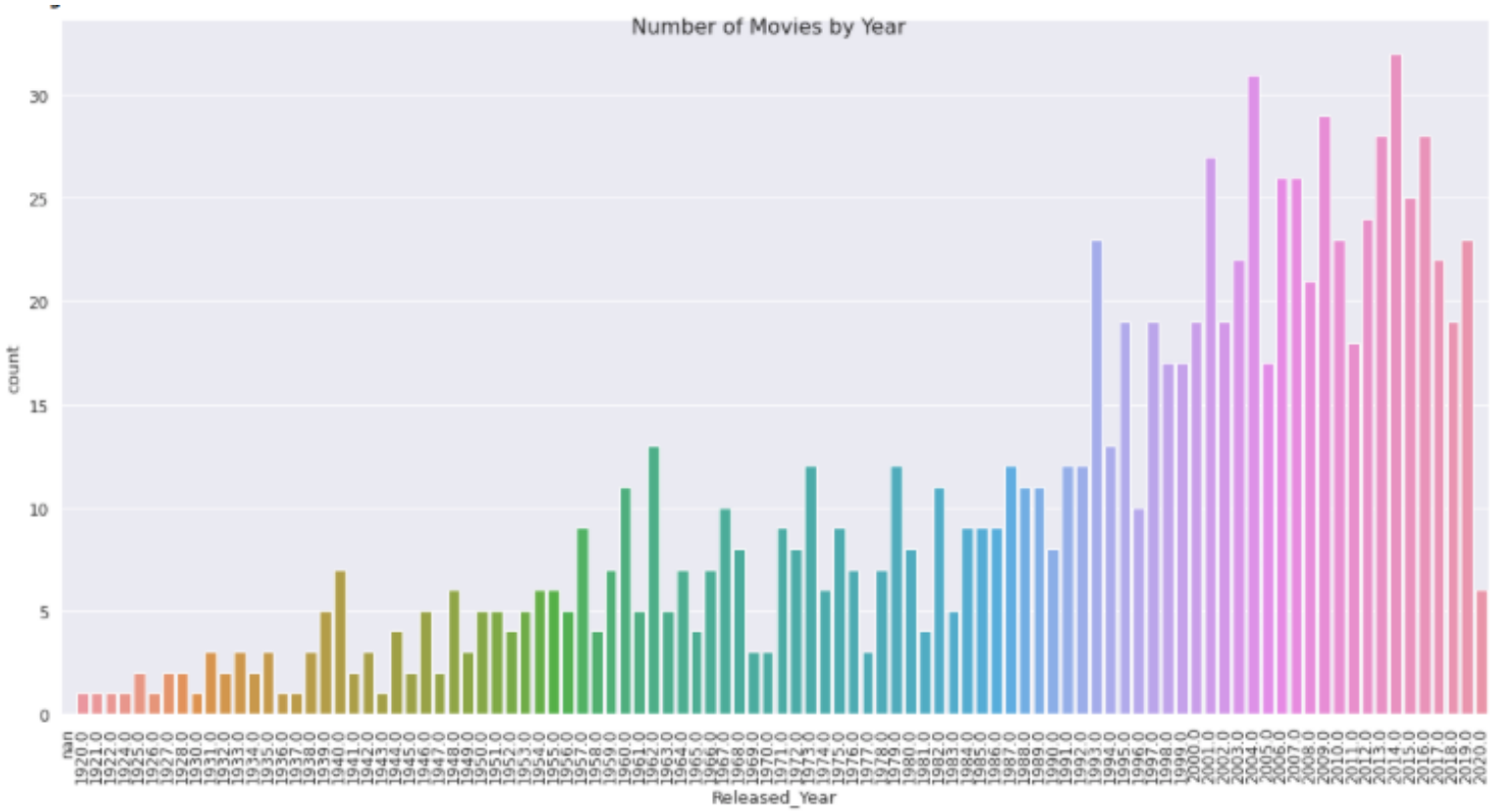
Dataset: The dataset contains each movie’s year of release, Meta Score, IMDb Score, Gross Revenue, and runtime. In terms of missing values, the dataset was missing 157 Meta Scores, most of which are for the really old movies. In order to quickly check over the data, I printed descriptive statistics for each column in the dataset. I counted the number of movies from each year, took the average Meta Score and IMDb Rating from each year. Click Here to view the data source. Click Here to view the source code.

Guiding Questions:

1. To what extent does recency of release influence the rankings?
2. Do the top ranked movies also have the best reviews?
3. What movies on the list have the greatest discrepancy between Meta Score and IMDb Rating? What might be the cause of this?

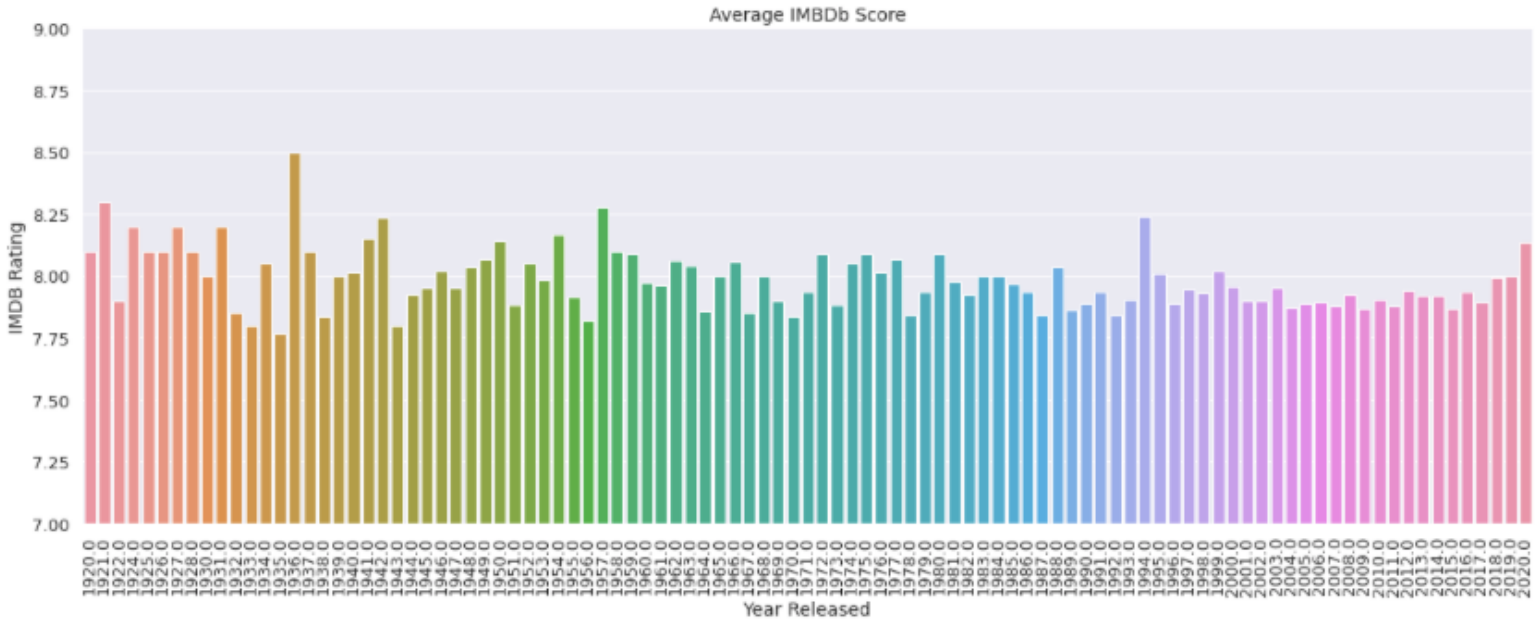
Number of Movies by Year

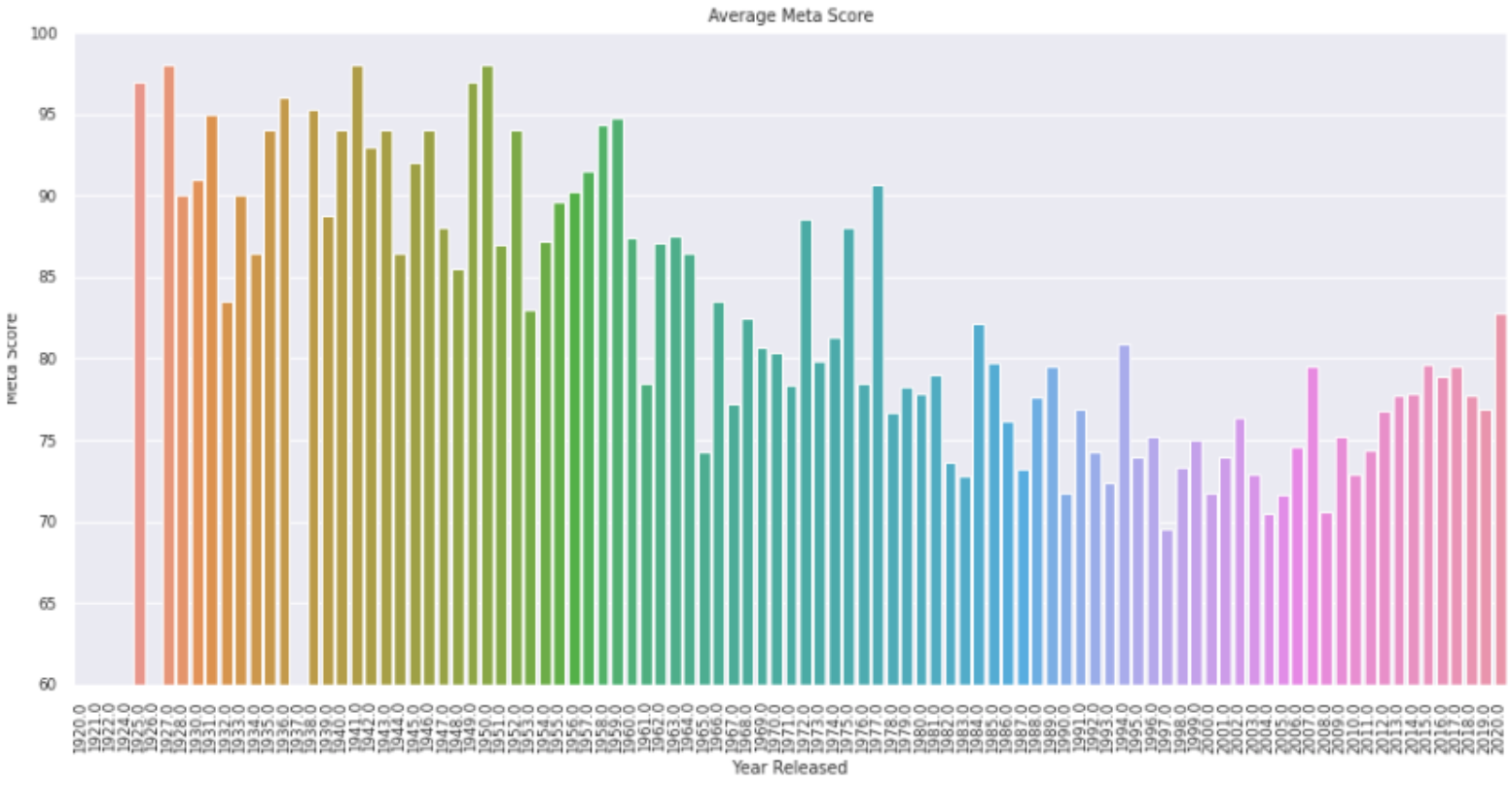
In order to get an idea of how recency effects the list, I counted the number of movies on the list from each year. The bar graph shows a steady increase over time from 1920 to 2020. The leading year being 2014.



Average IMDb Rating and Meta Score

While this was helpful to get an overall picture of the distribution of movies, I was curious as to how well the movies from each year were reviewed. The IMDb Rating is the key factor in determining these rankings which is why there is little difference between each year. The critics on the other hand, favor the older movies significantly. Thus, many movies that are considered in this list, have significantly poorer Meta Scores than their IMDb rating.


Average IMBDb Score
IMDB Rating
9.00
8.75
8.50
8.25
8.00
7.75
7.50
7.25
7.00
Year Released


Average Meta Score
Meta Score
100
95
90
85
80
75
70
65
60
Year Released

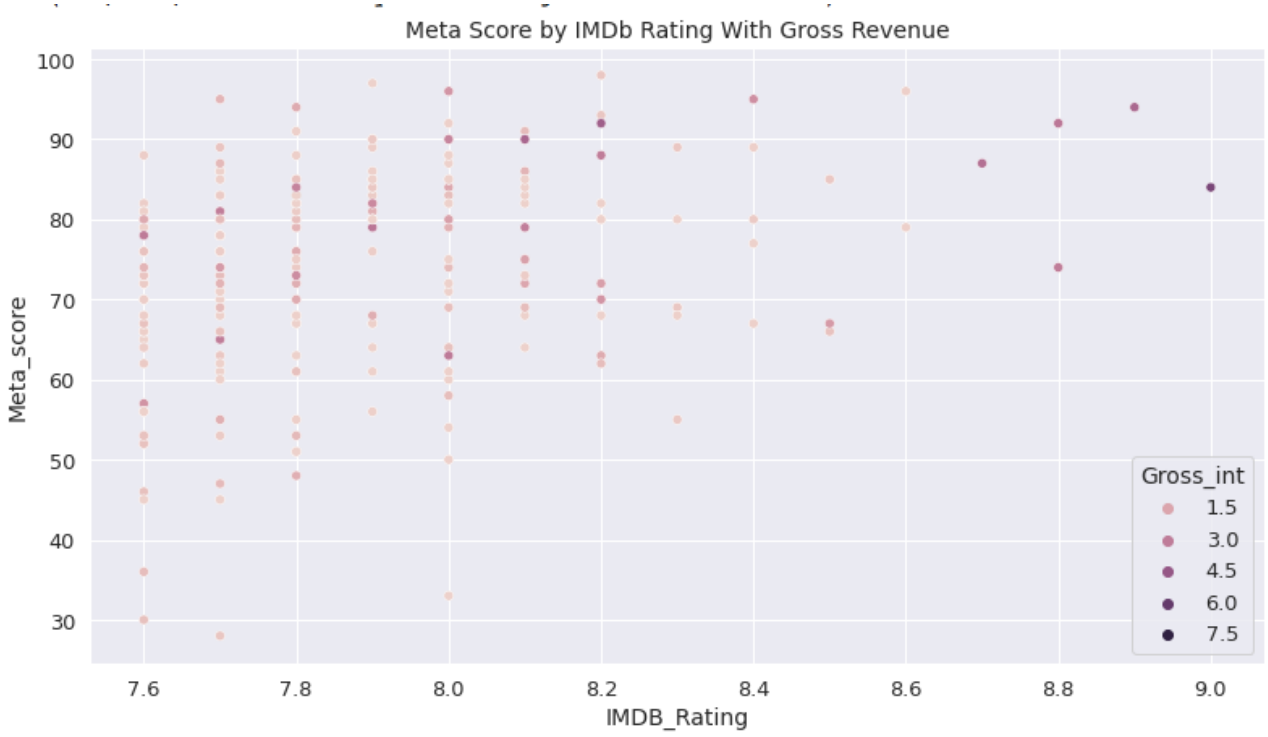
IMDb Rating (Popular) Versus Meta Score (Expert)

In order to compare the two different types of reviews, I had to convert IMDb rating to the same scale as the Meta Score rating by multiplying it by 100. Then by taking the difference and sorting it in descending order, I was able to see which movies the audience loved but the critics disliked and where they fell on the rankings (index).

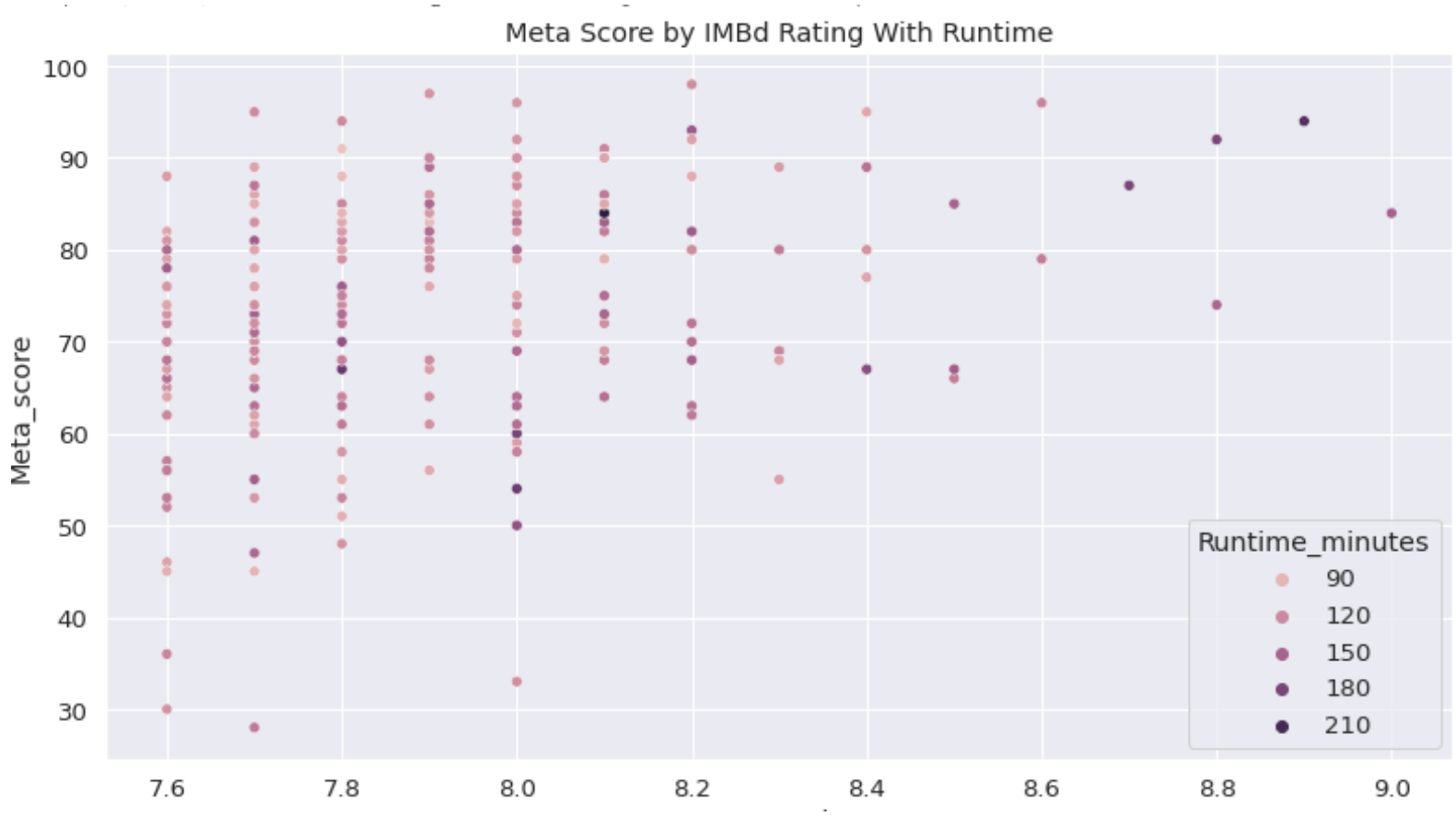
	Series_Title	Popular Versus Expert	Gross	Released_Year
788	I Am Sam	49.0	40,311,852	2001.0
356	Tropa de Elite	47.0	8,060	2007.0
942	The Butterfly Effect	46.0	57,938,693	2004.0
917	Seven Pounds	40.0	69,951,824	2008.0
735	Kai po che!	37.0	1,122,527	2013.0
957	Fear and Loathing in Las Vegas	35.0	10,680,275	1998.0
272	Pink Floyd: The Wall	34.0	22,244,207	1982.0
648	The Boondock Saints	34.0	25,812	1999.0
397	Bound by Honor	33.0	4,496,583	1993.0
677	Predator	33.0	59,735,548	1987.0
760	Flipped	32.0	1,752,214	2010.0
342	Bohemian Rhapsody	31.0	216,428,042	2018.0
935	Jeux d'enfants	31.0	548,707	2003.0
352	My Name Is Khan	30.0	4,018,695	2010.0
644	Remember the Titans	30.0	115,654,751	2000.0
782	Man on Fire	30.0	77,911,774	2004.0
811	Primal Fear	30.0	56,116,183	1996.0
932	Saw	30.0	56,000,369	2004.0
597	La migliore offerta	29.0	85,433	2013.0
35	The Intouchables	28.0	13,182,281	2011.0

IMDb Rating Versus Meta Score with Gross Revenue and Runtime

Using the other data given in the dataset, I was curious to see if there was any correlation with the movies that had high audience reviews with low critic reviews and the movies' gross revenue or runtime. There is a clear parallel between gross revenue and IMDb Rating and almost no correlation between Meta Score and the revenue the movie generated.

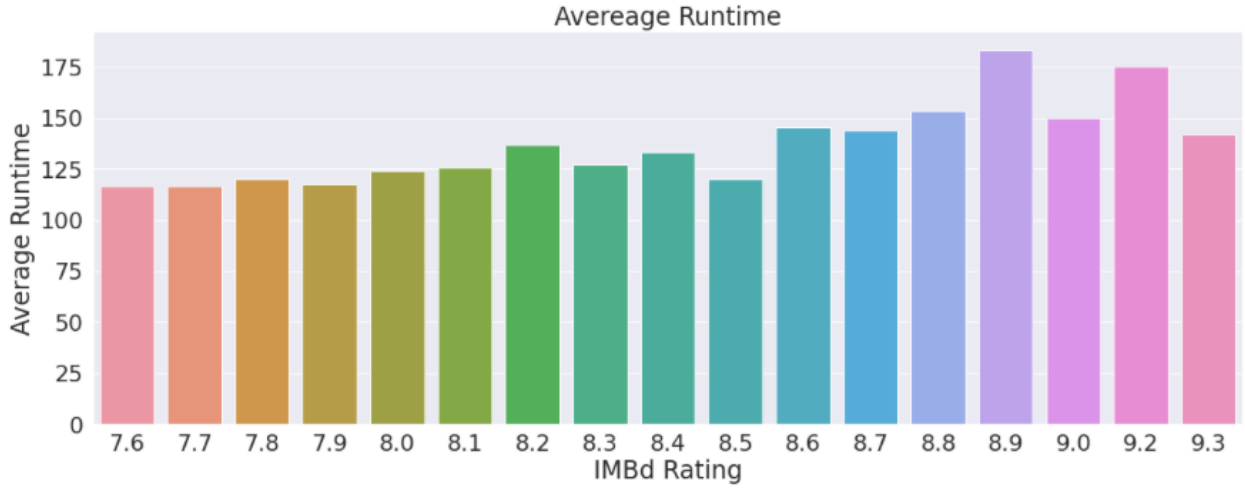


I thought part of the reason for a movie to get a high Meta Score and low IMDb rating might have to do with the length of the movie. This is because a movie critic is more likely to enjoy sitting through a long movie than an average person. However, there is no correlation with movie runtime and IMDb Rating or Meta Score.



Average Runtime per IMDb Rating

A surprising finding that I discovered is that the average runtime consistently increases as the IMDb score increases. This disputes my earlier prediction that the general population would prefer shorter movies. Although it is clear the greatest movies of all time tend to have longer run times.



Conclusion: According to the dataset, recency of release has some correlation with the Meta Score but not with the IMDb Rating. Since the rankings are built off the IMDb Rating, recency of release does not influence the rankings. Next, we saw little to no correlation between the Meta Score and the IMDb Rating. Therefore, the movies that the general public voted as the greatest movies of all time, are not considered to be the greatest movies of all time by professional movie critics. Lastly, there is an abundance of movies in this dataset that are considered bad movies by movie critics but not by the public. Interestingly, there is a much closer parallel to the amount of money a movie made with the IMDb rating as opposed to the Meta Score. Meaning that just because a movie gets bad reviews, does not necessarily mean it will do poorly in the box office. While the data in this dataset does not provide evidence for why this is the case, my guess is that movie critics and the general public simply look for different things in a movie.