**Homework #1  
Problem #1  
Timothy Tsai   
332003717**

*I certify that I have personally done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment*. TT.

# Part 1: Movies Movies Movies…

The goal of this problem is to conduct exploratory data analysis on a dataset of movie data to identify potential issues and uncover meaningful insights regarding the relationships between various movie characteristics. The process began by addressing missing values and ensuring the dataset was properly formatted, including converting relevant columns from text to numerical data. Once the data was cleaned, potential correlations between different movie attributes were visualized in a correlogram, shown in **Figure 1**, to reveal patterns and possible inconsistencies within the dataset.

A screenshot of a computer

Description automatically generated

Figure . A correlogram is utilized here to visualize different correlations between different movie attributes. Each pair of values being compared is given a score from -1 to 1, with numbers close to 1 meaning those elements are very strongly correlated

However, some of the results were unexpected. For instance, the correlation between "Year of Release" and "Metascore of movie" showed a negative relationship, which was surprising given the assumption that newer films would generally receive higher critical scores. Additionally, the correlation between "Movie Rating" and "Gross" was relatively weak, raising questions about how much movie ratings actually influence box office performance. These findings suggested that there are more complicated factors than ratings that relate to the success of movies.

The correlogram results as shown in **Figure 1** reveal several notable relationships between key movie attributes. One of the relationships was the negative correlation (-0.35) between the "Year of Release" and the "Metascore of movie." This suggests that newer movies, on average, tend to receive lower critical scores compared to older films. Additionally, another supporting factor for this assumption was the weak negative correlation between "Year of Release" and "Movie Rating" (-0.19) points to a potential decline in audience ratings for more recent films. These correlations suggest that both audience reception and movie critics alike have felt slightly less favorable to newer movies.

# Part 2: The Fall of Modern Movies

After the surprising results found out about the general dissatisfaction of newer movies, the next steps involved analyzing the average Metascore for each year over time. The aim of this was to clearly show if there was a negative correlation of movies scores through a time series graph. This was done by averaging the score of each movie in their respective year, then plotting those average over time shown by **Figure** **2**.

A graph showing a line going up

Description automatically generated with medium confidence

Figure . A time series graph shows the trend between the Year of Release (plotted on the x-axis) and the Average Metascore (plotted on the y-axis). Additionally, a regression line is drawn on top to show the relationship between these two attributes. Finally, the worst performing year is shown with its respective Metascore.

The results shown in **Figure 2** show the clear negative relationship shown in the time series, indicated by the downward trend of average Metascores over time. Given these results, a clear interpretation is that there is less appeal to newer movies. A potential interpretation of these results could be a general shift in the film industry’s need for commercially viable movies at the expense of quality. Additionally, a potential increase of movies over time could lead to the oversaturation of the market with a like of create content, which would lead to the general dissatisfaction of movie watchers.

# Part 3: Numerical Distributions of Movies

The objective of this analysis is to explore the distribution of various numerical attributes within the IMDB Movies Dataset, including 'Year of Release,' 'Watch Time,' 'Movie Rating,' 'Metascore of movie,' 'Gross,' and 'Votes' in **Figure 3.** By calculating the frequency of these attributes, insights into trends and patterns can be uncovered to better understand how movies stack up to each other.

A graph of a graph

Description automatically generated with medium confidence

Figure . Shown here is an array of histogram plots, each showing the frequency of different numerical attributes. Additionally, each histogram also has a kernel density estimation, which is the continuous probability density curve, for each numeric attribute.

The analysis of the distributions of numerical attributes in Figure 3reveals several significant trends. The 'Year of Release' distribution is left-skewed, indicating the dataset has more recent films, while 'Watch Time' follows a normal distribution, suggesting standard movie lengths perform better. The 'Movie Rating' is right-skewed, highlighting that many films receive average ratings given that this dataset contains the top 1000 movies, but few obtained very high scores. The 'Metascore of movie' is also lightly left-skewed, indicating critics also rated these movies fairly positively. In contrast, both 'Gross' and 'Votes' show heavy right skewness, suggesting that a small number of films dominate box office revenue and audience engagement. This very high skew likely points to very few films receiving the potential attention they should receive for how highly rated they are by critics. It could also mean that critics ratings play very little role in a film’s success.