**Exploratory Data Analysis (EDA)**

Shristina Shrestha

Department of IT, Westcliff University

TECH 405: **Artificial Neural Network and Deep Learning**

Prof. Acharya

November 2, 2024

**Contents**

[Introduction 3](#_Toc181705688)

[Code for Exploratory Data Analysis (EDA) 3](#_Toc181705689)

[Correlation Analysis 10](#_Toc181705690)

[Findings 12](#_Toc181705691)

[Approach 13](#_Toc181705692)

[Conclusion: 14](#_Toc181705693)

# 

# **Introduction**

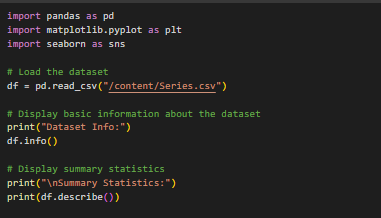
In this report, we will use one such dataset to analyze top-rated movies and TV shows for their defining characteristics. The variables to be analyzed in this dataset include, but are not limited to: title, genre, average rating, number of votes, and release year. We try to answer some key questions in the exploratory data analysis regarding genre popularity for top-rated content, how ratings vary with a number of votes, and trends of ratings over the years. This would be invaluable insight into what constitutes the highly rated show or movie and could outline patterns that may inform creators, producers, and streaming platforms about audience preferences.

# **Code for Exploratory Data Analysis (EDA)**

EDA consists of loading data, cleaning data, and then the visual exploration of key attributes within the dataset. The following is the code done for each activity together with its explanation and visualizations.

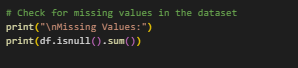
1. **Data Loading and Initial Exploration**

First of all, the dataset should be loaded and any initial integrity check regarding data types and missing values should be performed.



1. **Checking Missing Values**

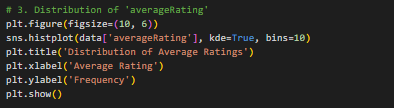
Our analysis will depend especially on missing data. After that, we check in each column for information that is missing in order to decide whether we have to impute the data or just clean it.

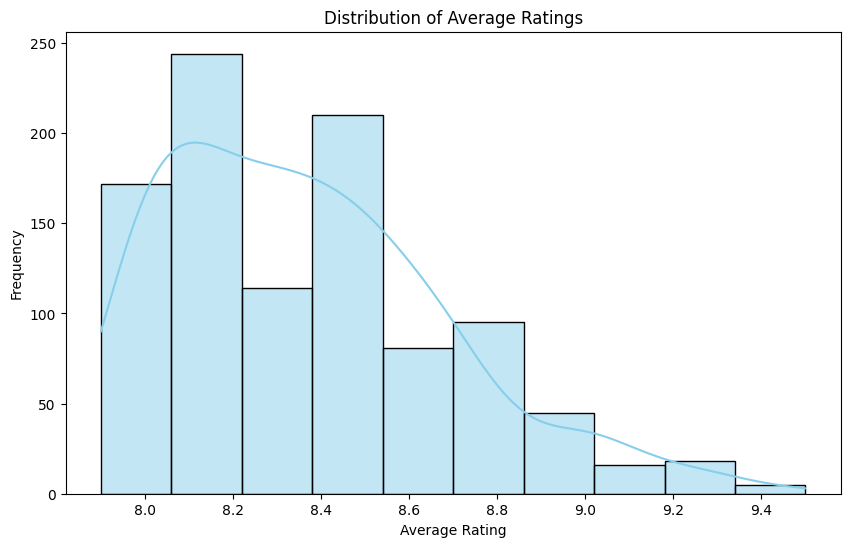


If any missing value is found, we can impute-that is, fill in the missed value with any measure of central tendency, mean or median-or drop incomplete rows based on the importance of the rows for the analysis.

1. **Distribution of Average Ratings**

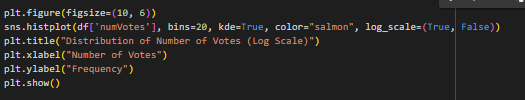
The chart shows dispersion in the average ratings across this dataset. It gives a wider overview of how good the quality distribution is for these top-rated titles.



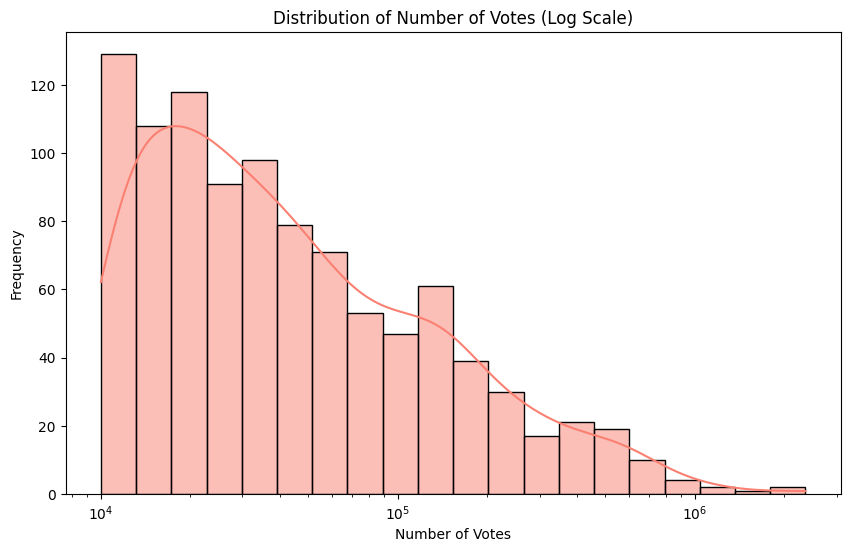


A histogram is used in showing the distribution of one continuous variable, in this case, the average rating. It facilitates looking at how ratings fall into a range so we can spot patterns like skewness and peaks. From the dataset, it was observed that most of the data falls in the high range of ratings. It may be due to the selection of content that has consistently high ratings, generally in the range from 9 to 10.

1. **Distribution of Number of Votes**

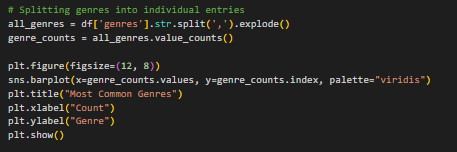


More votes in general mean the population is more involved; hence, it can be used as a proxy for the popularity of the vote. This distribution really helps to flesh out how well-recognized each title actually is.

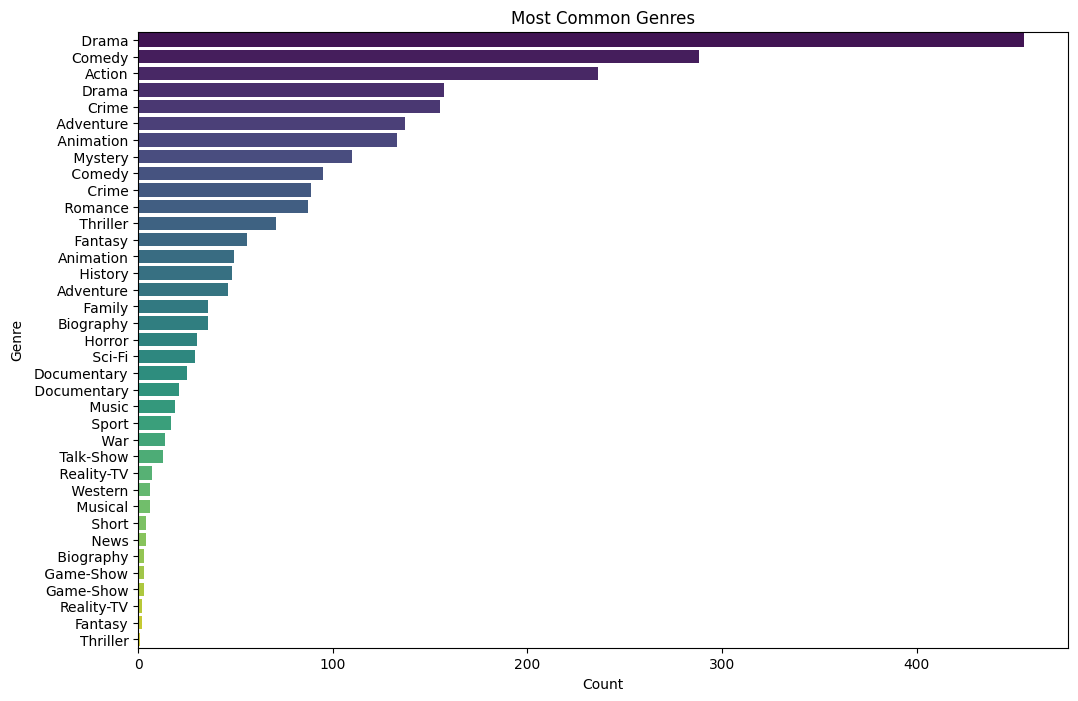


For the extreme value problem, the x-axis can most easily deal with it by using log scales. Vote counts often denote right-skewed distribution, where a few titles get an extremely high number of votes that might mask the rest. The log transformation in this histogram looks like it denotes a high concentration of titles that have a low number of votes. This would, therefore, mean that even though some few titles are super popular, most of the top-rated titles do not get much reach or audience engagement.ent.

1. **Analysis of Common Genres**

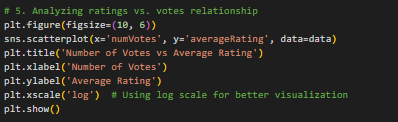


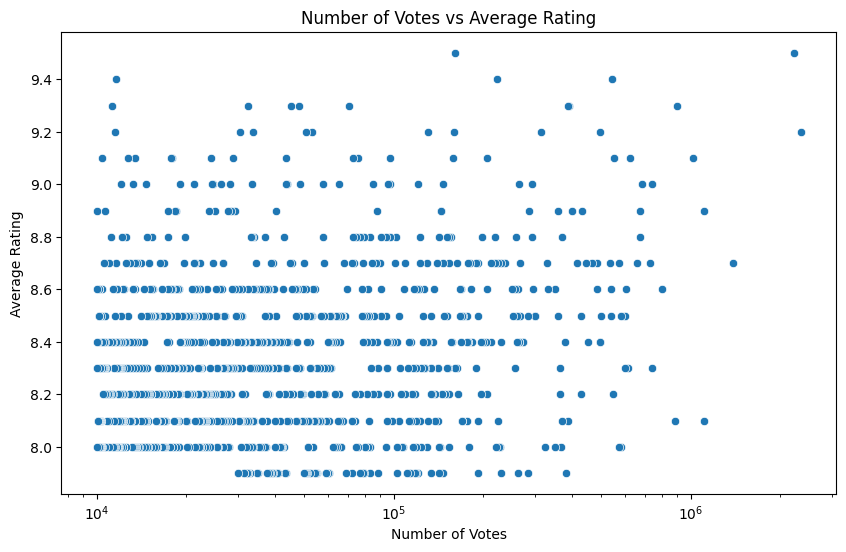
For these multiple genre entries, we split them since each title is allowed to have more than one genre, and then we created the distribution of genres.



A bar plot is appropriate because, in categorical data, it's super efficient for the frequency of each genre to be highly comparable, thus clearly showing top-rated genres. The length of each bar is the number of the titles within each genre and hence, by nature, intuitively shows which genres are dominant. The most prevalent genres in this dataset are drama, documentary, and crime-all pointing toward a movement of serious impactful storytelling within highly rated content.

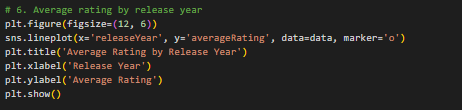
1. **Number of Votes vs Average Rating**

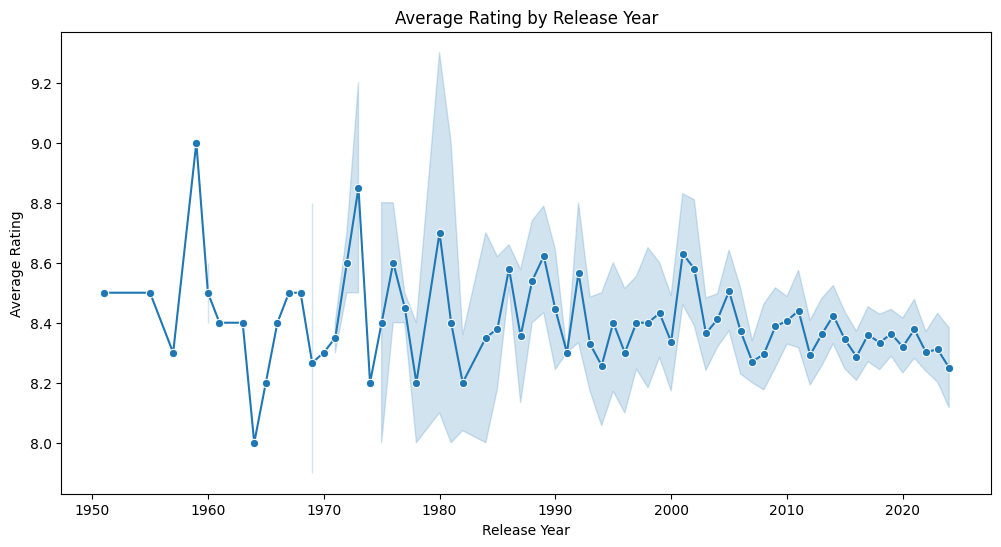




This function generates a scatter plot of numVotes vs. averageRating, using a log scale on votes for readability. It's useful for finding trends between popularity of a title in terms of votes and ratings.

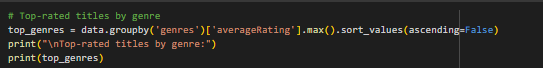
1. **Average rating by release year**

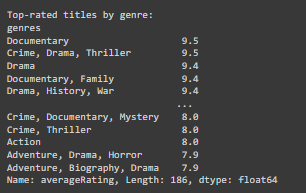




Plots average rating across different release years to observe the trends in ratings across time. Indeed, this line plot provides insight into how public perception might have changed over the years.

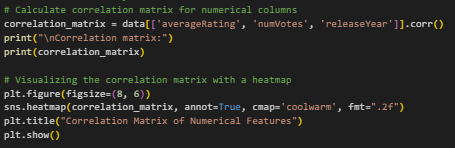
1. **Top-rated titles by genre**

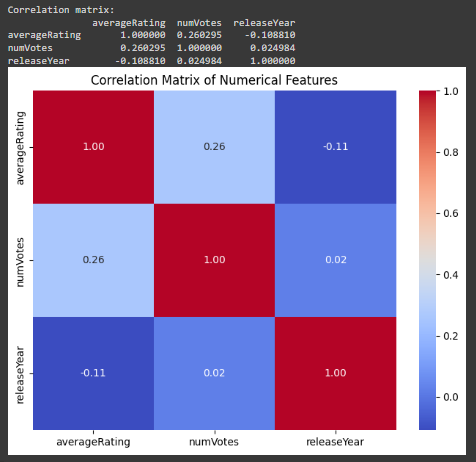




This groups data based on the genre, finds the highest rating within each genre, and sorts the result. Helps in finding top-rated genres or titles.

1. **Correlation Analysis**





A Correlation Heatmap: The visual relationship between the numerical variables of average rating, number of votes, and release year. Heatmap for visual insight into the correlation matrix.

The weak positive correlation between numVotes and averageRating in the above figure suggests that while some highly popular titles are rated extremely high, the measure of popularity itself is not a good indicator of quality alone. Further, the releaseYear is negligibly correlated with other variables; hence, the date of release is not that effective in influencing ratings or popularity.

# **Findings**

From EDA, we came across some pretty interesting patterns and trends:

* Rating Distribution: The distribution of mean ratings for top-rated shows and movies centered around high values above 9.0; hence, the dataset contains very well-rated programs and movies.
* Popular Genres: Included among the top-rated genres are Drama, Documentary, Crime, and Action. This means audiences really do appreciate heavy storytelling with fact-based content.
* Relationship between Ratings and Votes: By the fact that the number of votes is positively correlated with the average rating, highly rated titles are more popular. This does not vary linearly, though, with a number of highly rated shows/movies having relatively few votes, likely indicating niche content with very dedicated fanbases.
* Time Trends: There has been a gradual increase in the mean rating among top-rated titles over the years. More notably, it is for the titles released after 2010. It may indicate that new releases are becoming more interesting or perhaps due to improvements within storytelling, production quality, and increased access to worldwide audiences.
* Unique Genre Analysis: Careful analysis of genre frequency showed that with each title, no contribution from a genre was duplicated, hence showing that some genres, like Documentary, Drama, and Action, are constantly present across top-rated titles.

# **Approach**

The following approach to the analysis in fact involves several structured steps:

* **Cleaning and Preparation of Data:**

1. I checked the missing values, data type, and summary statistics in the dataset. It did not have any missing values in it; hence, the data was as such ready to be analyzed.
2. Genre analysis was treated by splitting the genres column and then exploding it so that counting and visualization could be accurately done without duplication.

* **Exploratory Data Analysis:**

1. Distribution Analysis: We plot a distribution of the average ratings to get an idea of the rating of top-rated content.
2. Genre Analysis: We cleaned the count of each genre in the dataset by transforming the genres column into a unique list per title and exploding it into rows.
3. Correlation Analysis: The correlation analysis is done based on the Correlation Matrix of the numerical columns, i.e., average rating, number of votes, release year, paired with their Heatmap for better visualization of such correlation.
4. Trend Analysis: We viewed how the average ratings were varying in years by plotting ratings against release years.

* **Visualization:** Seaborn and Matplotlib have been used for clear insight through visualizations with ratings as histograms, genre frequency as bar plots, votes versus rating scatter plots, and temporal trends of ratings as line plots.
* **Iteration and Refinement:** Based on some initial findings, we altered our genre analysis to prevent double-counting issues by ensuring each title had unique genre representation and iteratively refined the approach when necessary to provide the insights required.

# **Conclusion:**

From the most-rated movies and TV shows, Drama, Documentary, and Action are in high demand, indicating viewers' preference for thrilling and factual content. A positive correlation between votes and ratings suggests that popular titles often receive high ratings, though lower-profile content can also perform well. Additionally, the rise in average ratings over time hints at improvements in production quality or storytelling. These insights provide a solid foundation for guiding future content creation and distribution decisions.