

WQD7001 Principles of Data Science

Group 11

Group Project Part 1

|  |  |
| --- | --- |
| **Name** | **Matric Number** |
| Nasir Uddin Ahmed | S2015449 |
| Muhammad Shahzad Rafiq | S2150889 |
| Qixiang Yin | S2150692 |
| Low Boon Kiat | 17138399 |

Table of Content

[1. Team Members & Roles 1](#_Toc134479855)

[2. Project Abstract 1](#_Toc134479856)

[3. Introduction / Motivation 2](#_Toc134479857)

[4. Problem Statement 2](#_Toc134479858)

[5. Project Objective 3](#_Toc134479859)

[6. Scope & Domain 4](#_Toc134479860)

[7. Summarize Literature Review 4](#_Toc134479861)

[8. DS Pathway 5](#_Toc134479862)

[8.1 Data Set (Obtain & Clean) 6](#_Toc134479863)

[8.2 Collaborator/ End Users (Asking Questions) 10](#_Toc134479864)

[8.3 Exploratory Data Analysis (EDA) 11](#_Toc134479865)

[9. Conclusion 23](#_Toc134479866)

[10. References & Appendixes – Slide Presentation, Data, LR etc. 24](#_Toc134479867)

# Team Members & Roles

Team members of Group 11 are listed in the table below:

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Link to E-Portfolio** |
| Nasir Uddin Ahmed | Leader | <https://turjo7.github.io/PODSP/> |
| Muhammad Shahzad Rafiq | Maker | <https://drive.google.com/drive/folders/1N9omuJPiXIvuB1MaVVuNRxkmdKfbOxYT?usp=sharing> |
| Qixiang Yin | Oracle | <https://drive.google.com/drive/folders/1LRA72ofBMVxH0dk8bCLQilPMvQiY2llE> |
| Low Boon Kiat | Presenter | <https://drive.google.com/drive/folders/1pNAE1HyH5i0V6Sn8NTKYBdY2E7W7qwQL?usp=sharing> |

# Project Abstract

This project focuses on utilizing a dataset with attributes such as IMDb ID, title, plot, type, rated, year, release date, runtime, genre, director, writer, actors, language, country, awards, Metascore, IMDb rating, and IMDb votes. The objective is to leverage this dataset to develop a recommendation system for a streaming platform. The recommendation system aims to address the challenge of helping users discover relevant shows and movies by providing personalized suggestions based on their preferences. By analyzing user behavior, including watching history and interests, the system will generate accurate and tailored recommendations. The goal is to enhance user engagement and satisfaction on the platform. Key factors in the creation of the recommendation system include efficiency and scalability. The system will be built to quickly process and analyze the dataset, guaranteeing immediate recommendations with minimal wait times. Additionally, it will be scalable to accommodate an increasing user base and a rising content library. The performance of the recommendation system will be assessed using a number of metrics, including precision, recall, and F1-score. These metrics will evaluate the precision and applicability of the suggested readings, enabling the recommendation system to be adjusted and improved. The system's performance evaluation and subsequent deployment will improve user experiences and potentially boost the streaming platform's revenue.

# Introduction / Motivation

Streaming platforms like Netflix, Amazon Prime Video and Hulu are gaining popularity because they provide flexibility for users to watch their favorite TV shows and movies at any time on any device. These streaming services can attract young and modern users because of their wide variety of TV shows and movie content. It allows them to watch any missed program as per their availability. Due to the new culture of binge-watching TV shows and movies, users are consuming content at a fast pace. In 2019, Disney released its own streaming service Disney+, claiming it as an active competitor in the streaming media industry. A vast range of on-demand content are available on Disney+ including films, TV series and documentaries for many popular brands such as Disney, Pixar, Marvel, Star Wars and National Geographic. Currently, over 160 million users worldwide subscribe to Disney+.

On a streaming platform, very often there are thousands of shows to choose from. Helping users to find new shows they will appreciate is a challenge for streaming platform like Disney+. A recommendation system can overcome this challenge by proposing relevant shows to users based on their prior watching history and interests. For example, if a user watches a superhero movie, another superhero movie will be recommended to the user soon after. It helps users to find something they like on the streaming platform by offering personalized suggestions that the user is more likely to find interesting and relevant. The goal of a recommendation system is to enhance user experience through personalized recommendations and to keep users streaming the shows offered by the platform. From a business standpoint, the more relevant shows a user finds on the platform, the higher their engagement. This results in increased user retention and generates more revenue for the platform.

# Problem Statement

Based on a survey conducted by Statista in 2021 on thousands of streaming media users aged 18 to 60 in the United States, eight out of ten Americans who paid for online video content subscribed to Netflix at some point over the past twelve months. On the other hand, five out of ten Americans who took the survey subscribed to Disney+ (Richter, 2021). There is a gap of 30% difference between Netflix and Disney+ in terms of user base. However, considering that Disney+ was only launched in November 2019, it has huge potential to overtake Netflix as the world’s biggest streaming platform. In fact, Disney+ is aggressively forecasting that it will surpass Netflix subscriber numbers in 2024. To achieve this ambition, Disney+ has significantly increase their investment in developing high-quality content. On top of that, providing personalized user experience is a crucial aspect for the success of any streaming service provider. Suggesting the right content at the right time will boost user engagement and improve user retention rate, eventually leading to the growth in number of subscribers. Hence, developing an effective recommendation system is an important step for Disney+ in the pursuit of its ambition to the leading streaming platform in the near future.

Human are surprisingly bad at choosing between multiple options. We tend to get overwhelmed quickly and make poor choices when presented with many options. This phenomenon is known as paradox of choice (Schwartz, 2016). When the number of choices increases, the difficulty of knowing what is best increases. Having too many choices often led us to take more effort for making a decision and can eventually leave us feeling unsatisfied with our own choice. This phenomenon also applies to streaming media where users are presented with thousands of video content. According to consumer research, a typical streaming media user loses interest after spending 60 to 90 seconds reviewing 10 to 20 titles on one or two screens with chances of only 3 titles being reviewed in detail. At this point, the users will either find something of interest or the chances of them abandoning the streaming platform increases significantly (Gomez-Uribe & Hunt, 2015). This finding shows people hate choices because they are afraid of taking risks involved in decision making. This is where a recommendation system should be brought into play. It helps users to choose and makes the decision-making process easier for users.

# Project Objective

We formulated three project objectives as listed below:

1. To build a recommendation system for Disney+ shows and movies that is accurate, efficient and scalable.
2. To evaluate the performance of the recommendation system using a variety of metrics.
3. To deploy the recommendation system to production.

# Scope & Domain

The scope of this project is to build a recommendation system for Disney+ shows and movies. The domain of this project is application of data science in streaming media industry. Today, streaming media industry is rapidly growing. According to Market Research Report published by Fortune Business Insights in 2022, streaming media industry is projected to be worth US$330 billion by 2030. Personalizing user experience through recommendation systems has proven to help streaming media platform to grow the number of subscribers. Popular streaming service providers including Netflix, Amazon Prime Video, Hulu, Disney+, HBO Max, YouTube Premium and Apple TV+ have witnessed tremendous growth in their user base through integration of recommendation systems in their streaming platforms. In this project, we will focus on building a recommendation system for Disney+ shows and movies given it is a rising star in the streaming media industry.

# Summarize Literature Review

Disney+ is a well-known streaming service that offers access to a variety of content. Finding content that matches users' preferences might be difficult due to the site's enormous library of films, TV series and documentaries. Recommender systems and visualizations are two potential remedies for this issue.

Roy and Dutta (2022) conducted a systematic review of research on recommender systems. They classified several categories of recommender systems and examined their software, algorithms, datasets, simulation environments and performance indicators. Although recommender systems are helpful tools for filtering online information, the authors pointed out that they struggle with issues like scalability, cold-start and sparsity. The assessment also emphasized the need for additional study to create more effective and efficient recommender systems, especially when tackling bias and personalization issues. A functional prototype of a personalized real-time movie recommendation system that combines collaborative filtering and content-based filtering techniques was created by Zhang, J., Wang, Y., Yuan, Z., and Jin, Q. in 2020. Collaborative Filtering (CF) is a well-liked method for suggesting films. Traditional CF methods, however, have scaling problems. In contrast to conventional CF techniques, the new CF algorithm proposed in this research is faster and more scalable.

Ahuja et al. (2019) proposed a movie recommender system using K-means clustering and K-nearest neighbor. The creation of a movie recommender system employing machine learning methods like K-means clustering and K-nearest neighbor is presented in this research paper. The system's various components including its architecture, process flow, pseudocode, implementation and operation are all discussed. With an RMSE score of 1.081648, the results demonstrate that the proposed approach performs better than the current methodology. A large data technique is suggested by Awan, Khan, Nobanee, Yasin, Anwar, Naseem and Singh (2021) in their recommendation engine for forecasting movie ratings. Traditional recommendation systems have drawbacks such as the need for prior user history and habits in order to carry out the duty of recommendation. A hybrid recommendation system for films that makes use of the most effective ideas from collaborative filtering and content-based filtering as well as sentiment analysis of tweets from microblogging websites.

# DS Pathway

To complete a data science project successfully, there are numerous essential procedures that must be taken. First and foremost, it's crucial to specify the project's goal and the precise issue or query that has to be solved. This initial phase includes researching the domain context and conducting background research. The following step is data collecting, which entails locating, obtaining, and getting ready for analysis relevant data sources. This comprises cleaning, formatting, and preparation to assure the caliber and interoperability of the data. When the data is prepared, tools for exploratory data analysis are used to obtain insights and identify patterns or trends. To completely comprehend the properties of the data at this level, visualizations and statistical techniques are frequently used.

The proper machine learning algorithms or statistical models are chosen and implemented to create predictive or descriptive models based on exploratory investigation. Once the model has been validated, an evaluation of its performance and generalizability is conducted. The last phase is to effectively communicate the outcomes and conclusions to the stakeholders via visualizations, reports, or presentations. This makes sure that the data science project's insights can be applied to decision-making and produce results that can be put into practice. To sustain ethical and responsible data practices, data ethics, privacy, and security issues should be carefully addressed across the entire pathway.

## Data Set (Obtain & Clean)

Obtaining Dataset: The Disney+ Shows Dataset is a valuable resource for researchers and practitioners in the field of data analysis and recommendation systems. This dataset contains information about various shows available on the Disney+ streaming platform, including attributes such as IMDb ID, title, plot, type, rated, year, release date, runtime, genre, director, writer, actors, language, country, awards, Metascore, IMDB rating, and IMDB votes.

*Table 1: Dataset Information*

|  |  |
| --- | --- |
| **Property** | **Value** |
| Data Source | <https://www.kaggle.com/datasets/unanimad/disney-plus-shows> |
| Data Name | Disney Plus Movies and TV Shows |
| Data Size | 408 KB |
| Year | 2020 |
| Dimension | 992 Rows & 19 Columns |

Cleaning the Dataset: A crucial phase in the data analysis process is dataset cleansing, commonly referred to as data cleaning or data pretreatment. Assure data quality and reliability, it entails locating and fixing flaws, inconsistencies, and inaccuracies in the dataset. Tasks include addressing missing values, eliminating duplicate entries, standardizing data formats, and resolving inconsistencies in variables or characteristics are examples of cleaning the dataset. Preparing data for analysis or modeling may also require converting and normalizing it.

*Table 2: Summary of Statistics of Each Attribute in the Dataset*

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| imdb\_id | object | Unique identifier for each IMDb entry |
| title | object | Title of the show or movie |
| plot | object | Summary or description of the show or movie |
| rated | object | Rating of the content (e.g., PG-13, R) |
| year | object | Year of release |
| released\_at | object | Date of release |
| added\_at | object | Date when the content was added to Disney+ |
| runtime | object | Duration of the content |
| genre | object | Genre or genres of the content |
| director | object | Director(s) of the content |
| writer | object | Writer(s) of the content |
| actors | object | Actors or cast members in the content |
| language | object | Language(s) of the content |
| country | object | Country or countries of production |
| awards | object | Awards or accolades received by the content |
| metascore | float64 | Metascore rating assigned by critics (0-100 scale) |
| imdb\_rating | float64 | IMDb rating given by users (0-10 scale) |
| imdb\_votes | object | Number of votes received on IMDb |
| type | object | Type of content (e.g., TV show, movie) |

Initially the dataset contained 992 rows and 19 columns. Among the 19 columns or attributes 17 of them were categorical and 2 of them were numerical.

For cleaning the data we have done several things they are:

1. Checking for the Missing Values: Checking for missing values is a crucial step in data preprocessing and cleaning. Missing values occur when data is not available or not recorded for certain observations or attributes in a dataset.

The table represents the missing values of each of the attributes of our dataset.

*Table 3: Summary of the Missing Values of the Attributes*

|  |  |
| --- | --- |
| **Attribute Name** | **Number of Missing Values** |
| imdb\_id | 98 |
| title | 98 |
| plot | 126 |
| type | 98 |
| rated | 250 |
| year | 98 |
| released\_at | 118 |
| runtime | 154 |
| genre | 107 |
| director | 303 |
| writer | 249 |
| actors | 122 |
| language | 127 |
| country | 123 |
| awards | 436 |
| metascore | 700 |
| imdb\_rating | 113 |
| imdb\_votes | 113 |

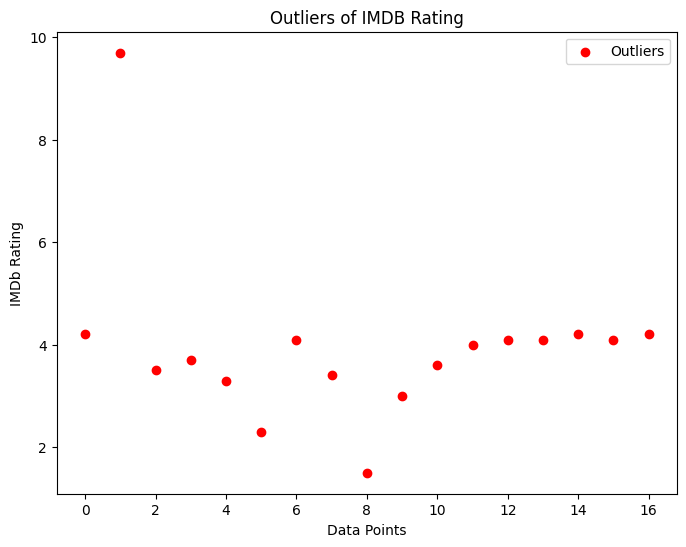
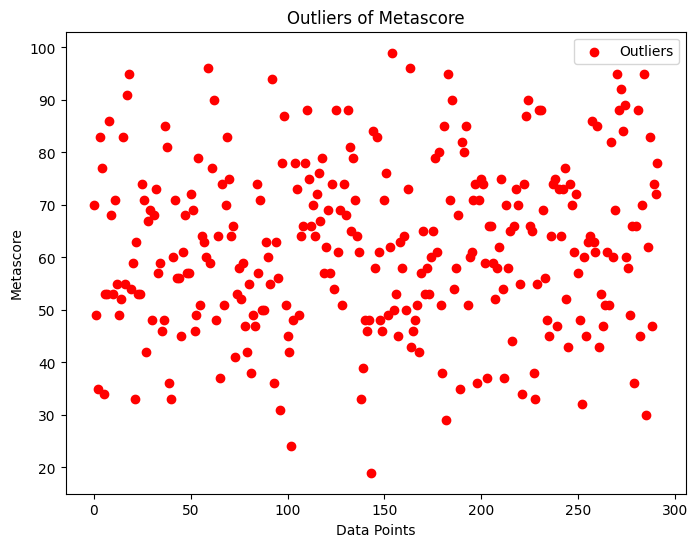
For the numerical features, we used mean and median to fill up the missing values. For "metascore" the distribution was approximately symmetric and not heavily skewed. For “imdb\_rating” the distribution of the variable with missing values was skewed.

For handling the missing values in categorical features, we used the most frequent value, as an imputation method.

ii. Checking for Duplicate Entries: Checking for duplicate entries in a dataset is an important step to ensure data quality and avoid potential issues in analysis or modeling. We found 74 observations had duplicate entries. We dropped those 74 rows from our data frame.

iii. Checking for Outliers: Outliers are typically identified and dealt with in numerical data, as the concept of outliers is not directly applicable to categorical data. Categorical variables represent distinct categories or groups, and outliers are more relevant in the context of continuous or numerical variables.

*Figure 1: Plot Showing the Outliers of Numerical Features*

The above figures show the outliers of our numerical features. We applied log transformation to handle the outliers. But log transformation shortens the range and distribution of imdb\_rating. So we applied another technique which is Winsorization. Winsorization replaces extreme values with the nearest values within a specified range.

## Collaborator/ End Users (Asking Questions)

Collaborators:

Data scientists and analysts: They will contribute their expertise in data analysis, machine learning, and recommendation system development to the project.

Streaming platform experts: Collaborating with professionals familiar with the streaming media industry will provide valuable insights into user preferences, content categorization, and platform requirements.

Disney+ stakeholders: Engaging with representatives from Disney+ will ensure alignment with their vision, goals, and platform-specific considerations.

End Users (Asking Questions):

How will the recommendation system personalize suggestions based on user preferences?

Can the recommendation system consider factors like genre, director, or actors in providing tailored recommendations?

What measures will be taken to ensure the recommendation system's efficiency and scalability as the user base and content library grow?

How will the system handle user privacy and data protection concerns?

Will the recommendation system provide explanations or justifications for the suggested shows and movies?

How will the performance of the recommendation system be evaluated and improved over time?

Can the recommendation system adapt to evolving user preferences and changing trends in the streaming media industry?

What steps will be taken to ensure the accuracy and relevance of the recommendations?

Will the recommendation system consider collaborative filtering, content-based filtering, or a hybrid approach in generating suggestions?

How will the deployment of the recommendation system impact the user experience on the Disney+ platform?

## Exploratory Data Analysis (EDA)

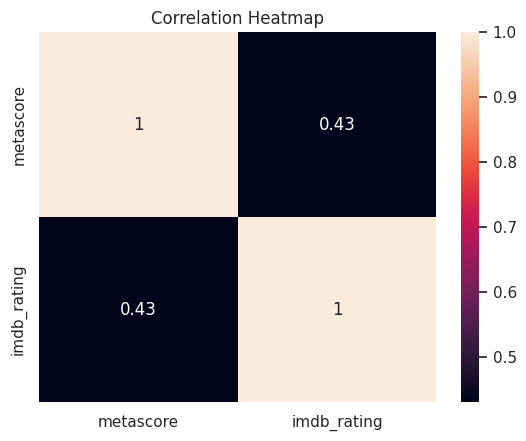
To gain insights, recognize patterns, and extract useful information from a dataset, exploratory data analysis (EDA), also known as exploratory data mining (EDM), is a critical first stage in the data analysis process. Researchers and data analysts can investigate and comprehend the traits, connections, and distributions of the variables in the dataset with the aid of statistical methodologies, visualizations, and data summarization approaches. EDA offers useful insights that guide further data preparation, feature engineering, and modelling procedures by analyzing the data, recognizing outliers, missing values, and comprehending the data's structure. EDA is a crucial tool in data-driven decision making since it not only helps to reveal hidden patterns and trends but also helps to formulate insightful research questions and hypotheses.

Descriptive statistics is a branch of statistics that involves the analysis and summary of data to provide meaningful insights and a concise overview of its main characteristics. It focuses on measures that describe the central tendency, variability, and distribution of the data. The following table is showing the descriptive statistics of the numerical variables of our dataset.

*Table 4: Descriptive Statistics of the Numerical Variables*

|  |  |  |
| --- | --- | --- |
| Statistics | metascore | imdb\_rating |
| count | 918 | 918 |
| mean | 62.06 | 6.65 |
| std | 8.88 | 0.99 |
| min | 19.00 | 1.50 |
| 25% | 62.06 | 6.10 |
| 50% | 62.06 | 6.70 |
| 75% | 62.06 | 7.30 |
| max | 99.00 | 9.70 |

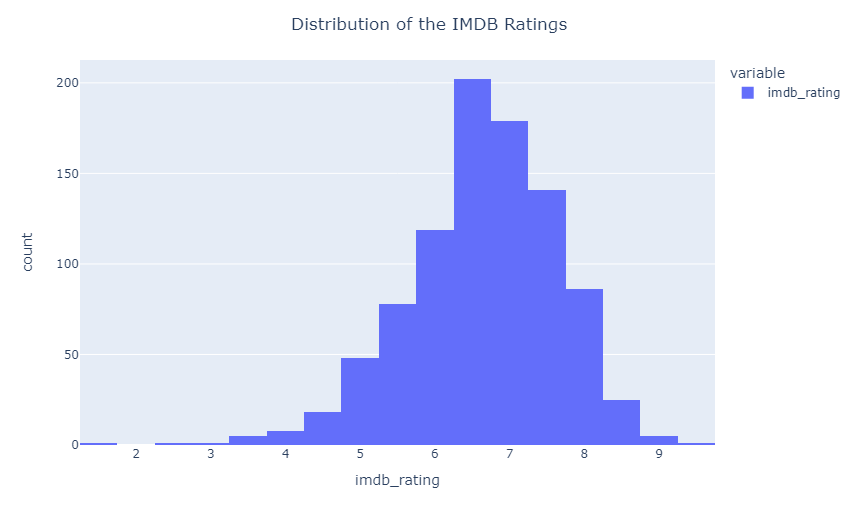
*Figure 2: Correlation Heatmap*



The correlation table shows the correlation coefficients between 'metascore' and 'imdb\_rating'. The correlation coefficient of 1.0 represents the correlation of a variable with itself, which is always 1.

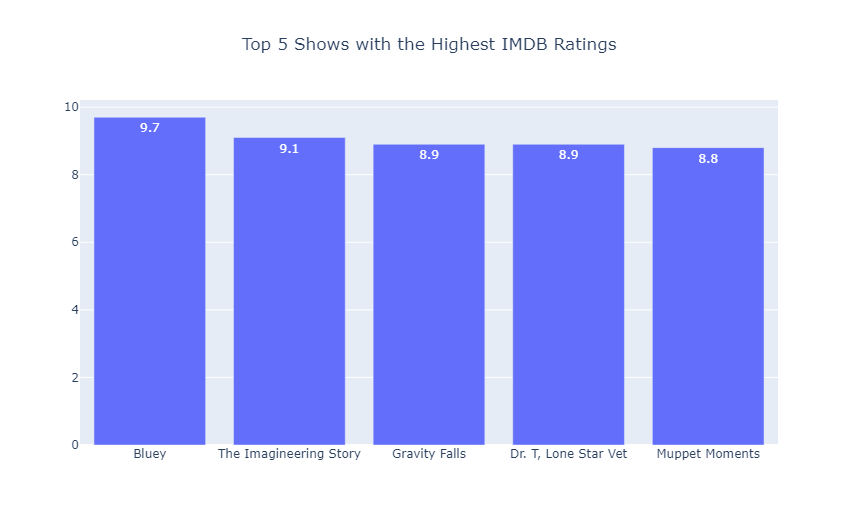
The correlation coefficient between 'metascore' and 'imdb\_rating' is 0.43. This indicates a positive correlation between the two variables, suggesting that there is a tendency for higher 'metascore' values to be associated with higher 'imdb\_rating' values, and vice versa. However, the correlation is not very strong, indicating that the relationship between the two variables is moderate.

*Figure 3: Distribution of IMDB Ratings*



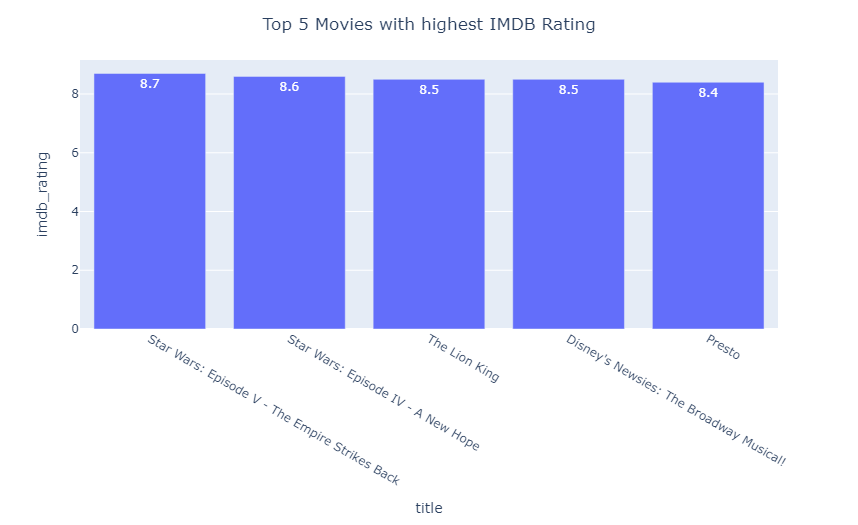
A histogram of the IMDB ratings in the Data Frame df. The histogram has 40 bins and the values are labeled as imdb\_rating. The histogram's title is "Distribution of the IMDB Ratings" centered in the above figure.

*Figure 4: Top 5 Shows with Highest IMDB Ratings*



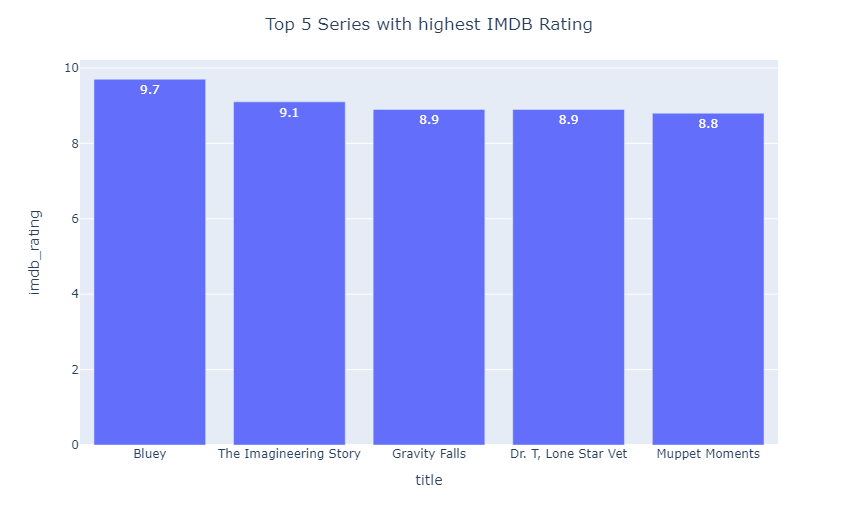
A bar chart of the top 5 shows the highest IMDB ratings. The bar chart shows the title of the show, the IMDB rating, and the text of the IMDB rating. The bar chart's title is "Top 5 Shows with the Highest IMDB Ratings" centered in the above figure.

*Figure 5: Top 5 Movies with Highest IMDB Ratings*



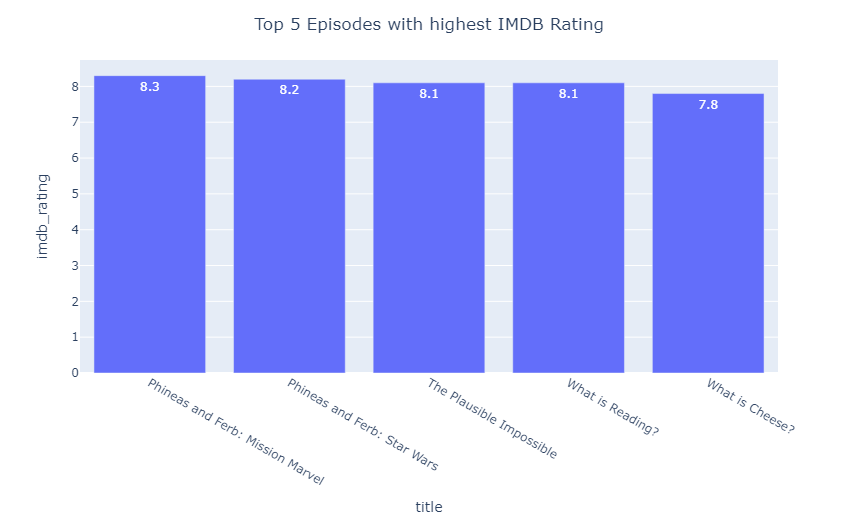
A bar chart of the top 5 movies the highest IMDB ratings. The bar chart shows the title of the movies, the IMDB rating, and the text of the IMDB rating. The bar chart's title is "Top 5 Movies with the Highest IMDB Ratings" centered in the above figure.

*Figure 6: Top 5 Series with Highest IMDB Ratings*



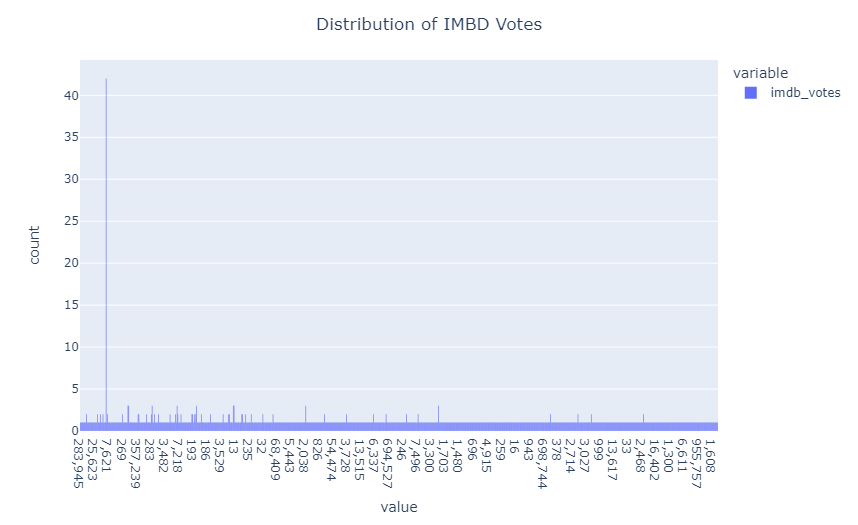
A bar chart of the top 5 Series the highest IMDB ratings. The bar chart shows the title of the series, the IMDB rating, and the text of the IMDB rating. The bar chart's title is "Top 5 Series with the Highest IMDB Ratings" centered in the above figure.

*Figure 7: Top 5 Episodes with Highest IMDB Ratings*



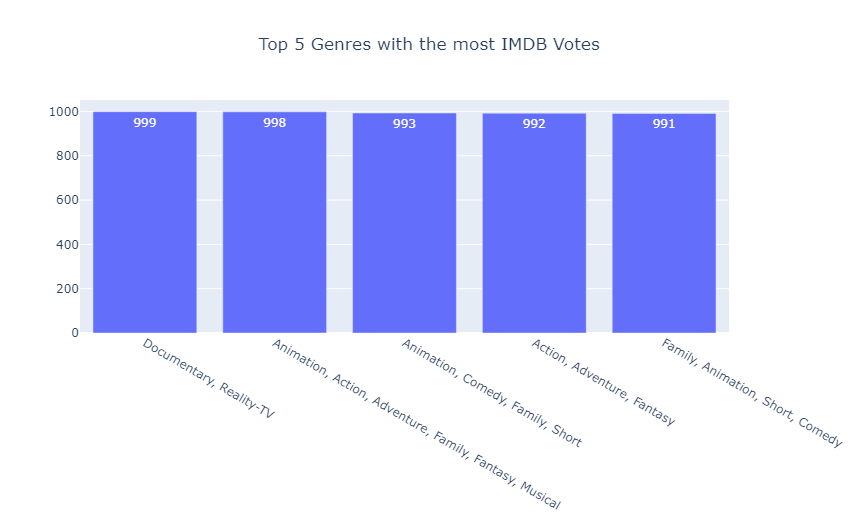
A bar chart of the top 5 episodes the highest IMDB ratings. The bar chart shows the title of the episodes, the IMDB rating, and the text of the IMDB rating. The bar chart's title is "Top 5 Episodes with the Highest IMDB Ratings" centered in the above figure.

*Figure 8: Distribution of IMDB Votes*



The chart above shows the distribution of the IMDB votes. The plot of the distribution of IMDb votes provides valuable insights into the popularity and engagement of movies or TV shows within the IMDb community. It typically displays the number of titles on the x-axis and the corresponding frequency or count of IMDB votes on the y-axis.

*Figure 9: Top 5 Genres with the most IMDB Votes*



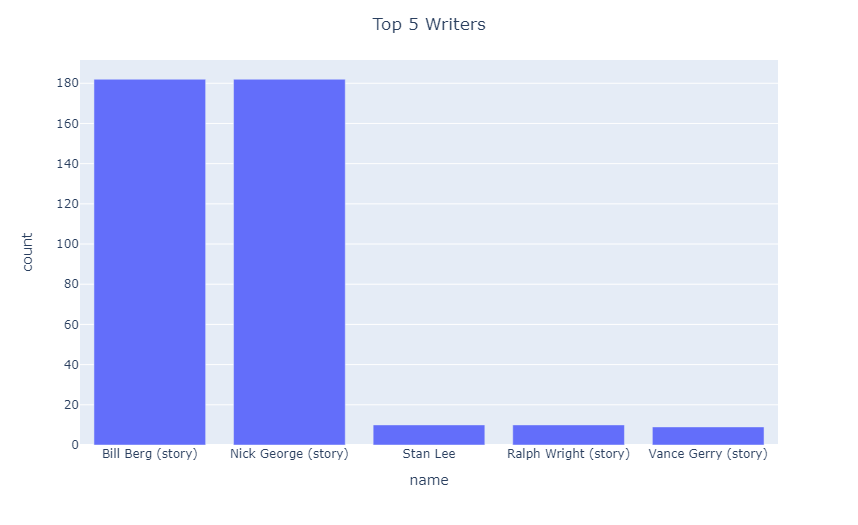
A bar chart of the top 5 genres with the highest IMDB ratings. The bar chart shows the title of the genres, the IMDB rating, and the text of the IMDB rating. The bar chart's title is "Top 5 Genres with the Highest IMDB Ratings" centered in the above figure.

*Figure 10: Word Cloud with the Most Frequently Words in the Title*



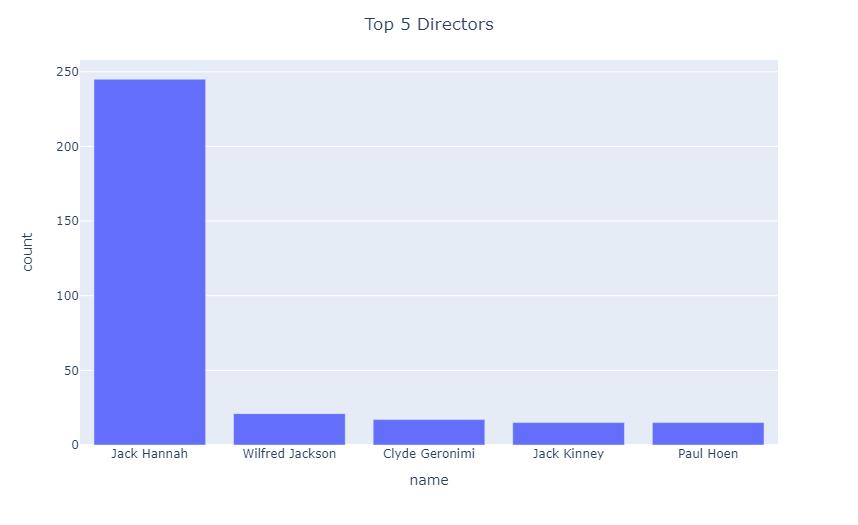
The word cloud shows the most frequent words used in the title.

*Figure 11: Top 5 Writers*



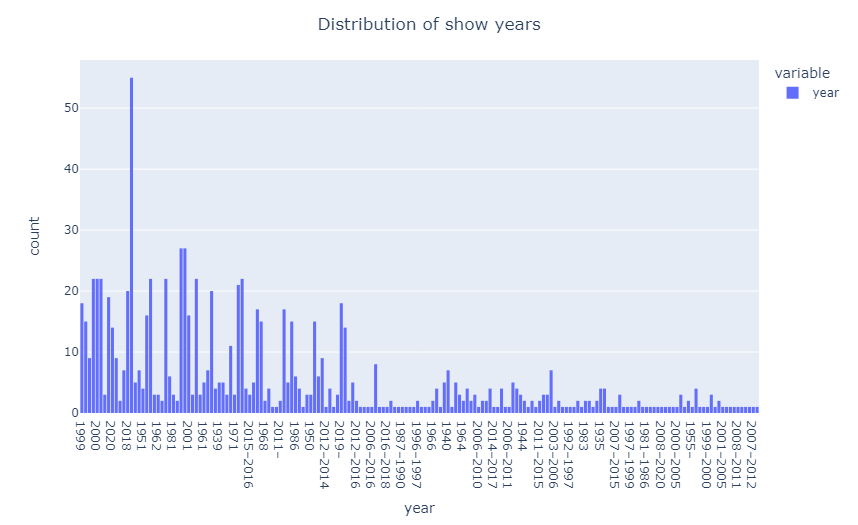
Bill Berg and Nick George writing story books over 180 however Stan Lee, Ralph Wright and Vance Gerry not enough 20 books, by the way Stan Lee is a top 5 writers only didn’t write story books writer.

*Figure 12: Top 5 Directors*



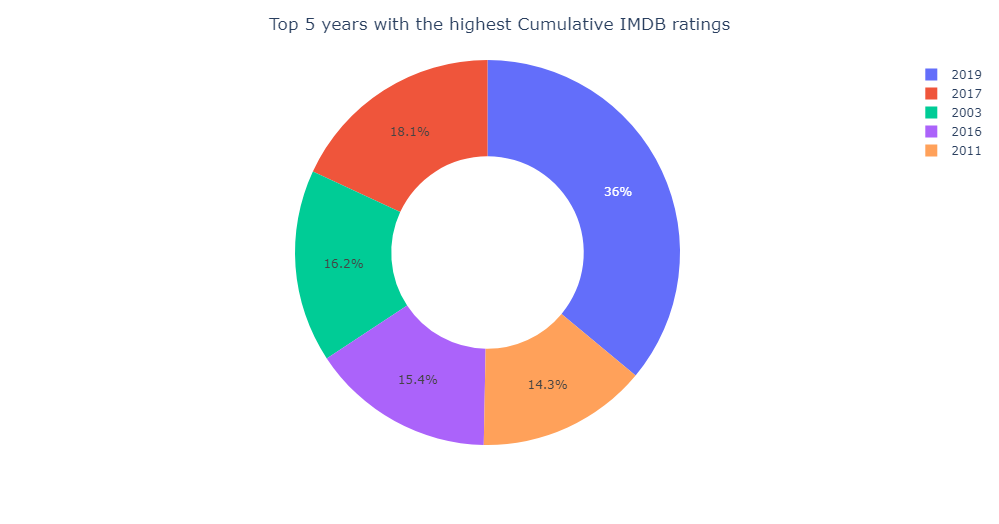
In the Top 5 directors only Jack Hannah filming almost 250 movies, other directors like Wilfred Jackson, Clyde Geronimi, Jack Kinney and Paul Hoen not enough 25 movies.

*Figure 13: Distribution of Shows Per Year*



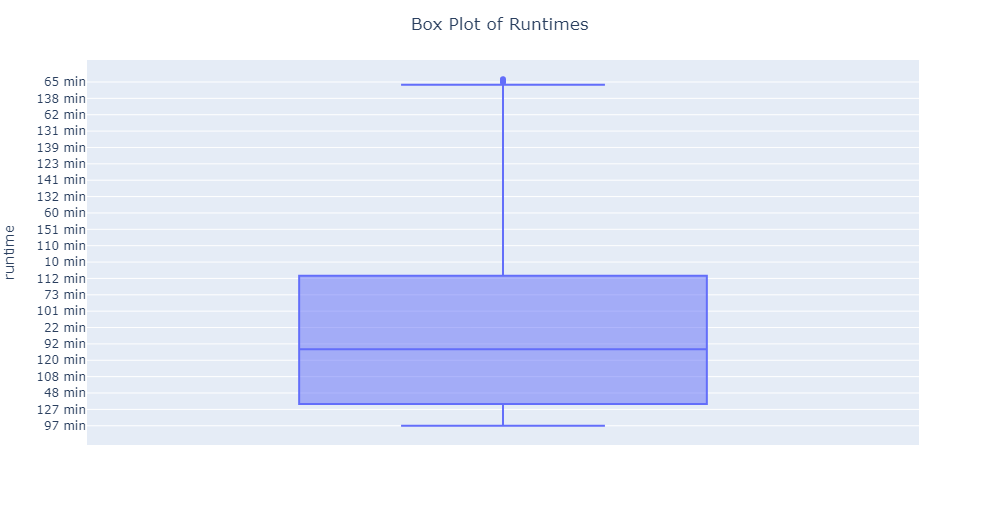
The plot of the distribution of shows per year provides an overview of the number of shows released in different years. This type of plot is commonly represented using a bar chart or a line graph, with the x-axis representing the years and the y-axis displaying the count or frequency of shows.

*Figure 14: Top 5 years with the Highest Cumulative IMDB Ratings*



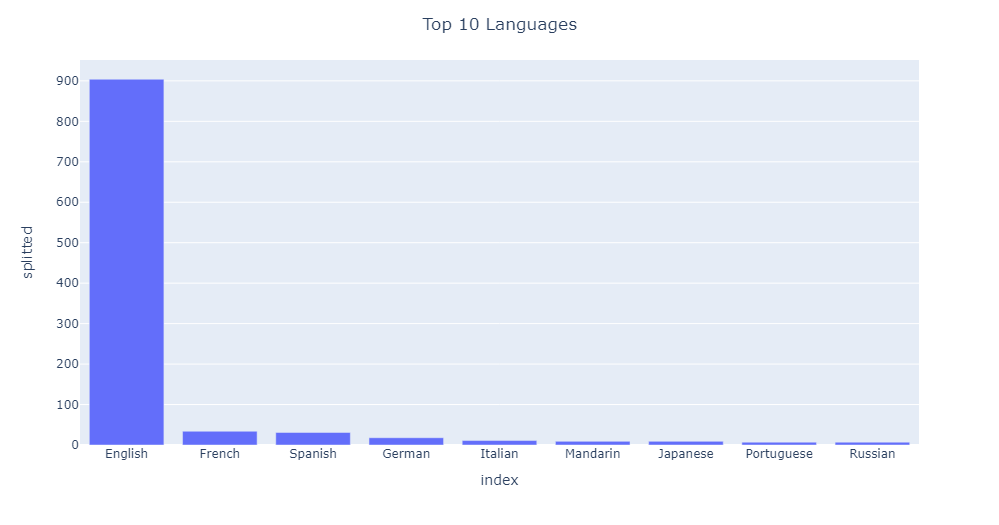
In 2019, IMDB has the highest cumulative ratings. It is taking 36 percent within the most early in 2003 occupied 16.2 percent higher than 2011 14.3 percent and 2016 15.4 percent, besides that, in 2017 have rapidly increased to the 18.1 percent.

*Figure 15: Box Plot of Runtimes*



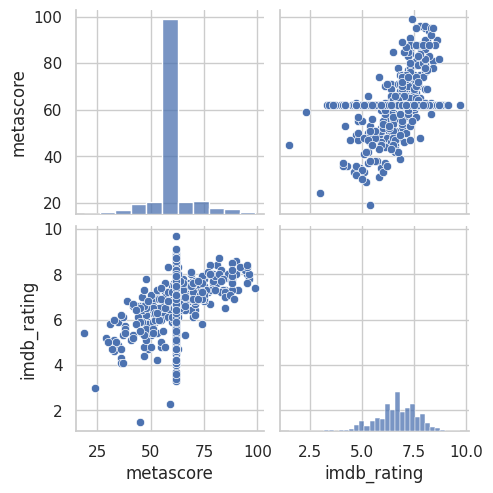
The box plot of run time helps identify the central tendency and spread of the run time values. The position and length of the box indicate the range of durations in which the majority of movies or TV shows fall. A longer box signifies a wider spread of run times, while a shorter box indicates a narrower range. The median line provides insight into the typical or average run time.

*Figure 16: Top 5 Languages*



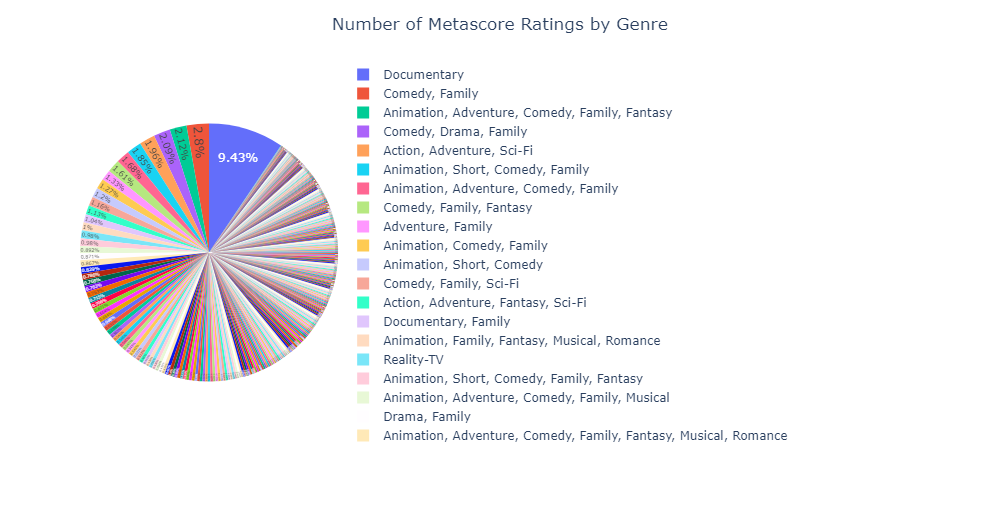
Most of language are using English, which is over 900, other languages usage rate had huge gap with English, following by French, Spanish, German, Italian, Mandarin, Japanese, Portuguese and Russian.

*Figure 17: Pair Plot*



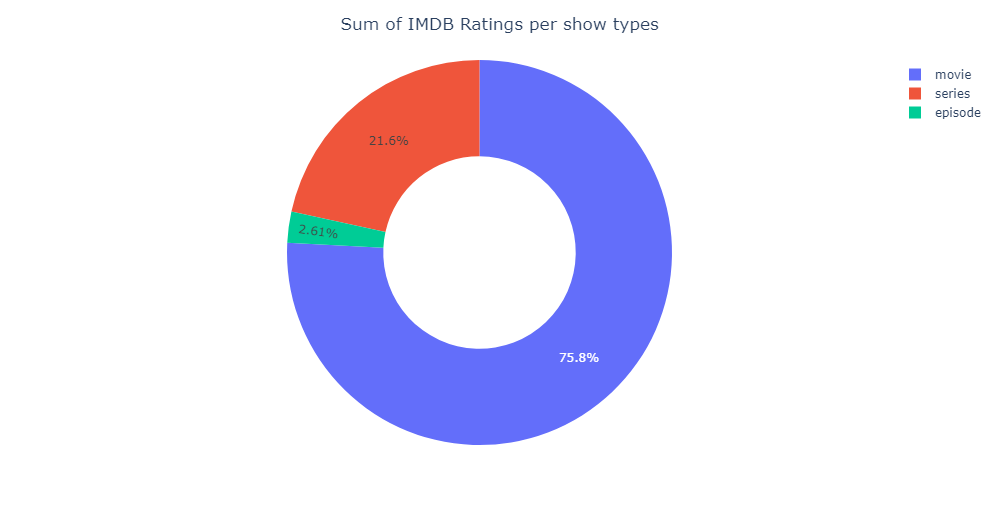
A pair plot of metascore and imdb\_rating provides a comprehensive visual representation of the relationship between these two popular rating systems for movies. It typically consists of a grid of scatter plots, where each plot displays the metascore on one axis and the IMDB rating on the other axis.

*Figure 18: Number of Meta Score Ratings by Genre*



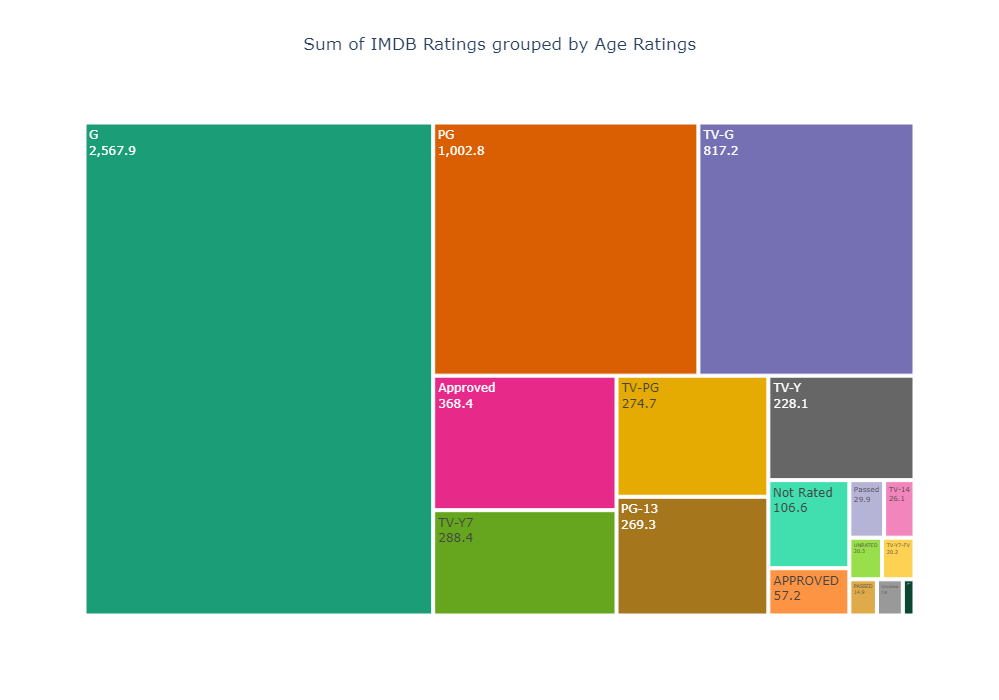
A plot displaying the number of metascore ratings by genre provides an informative visual representation of how many movies in each genre have been evaluated and assigned a metascore rating.

*Figure 19: Sum of IMDB Ratings per Show Types*



In IMDB ratings, movies are occupation large share which is around 75.8 percent. After movies is series, it takes 21.6 percent. Other few parts of IMDB ratings are taken by episode is only 2.61 percent.

*Figure 20: Sum of IMDB Ratings Grouped by Age Ratings*



A tree map of the IMDB ratings grouped by age ratings. The tree map shows the age rating, the sum of the IMDB ratings for that age rating, and the text of the sum of the IMDB ratings. The tree map's title is "Sum of IMDB Ratings grouped by Age Ratings" centered in the above figure.

# Conclusion

This project aims to build a recommendation system for Disney+ shows and movies, leveraging data science techniques to personalize the user experience. The project's scope is within the rapidly growing streaming media industry, where recommendation systems have proven to be effective in attracting and retaining subscribers. By focusing on Disney+, a rising star in the industry, the project aims to capitalize on its potential for growth. Exploratory data analysis (EDA) is a critical step in understanding the dataset and extracting useful insights. It involves investigating and analyzing the dataset's variables, patterns, and distributions using statistical methodologies, visualizations, and data summarization approaches. EDA helps to uncover hidden patterns and trends, identify outliers and missing values, and guide further data preparation and modeling procedures.

Descriptive statistics, a branch of statistics, play a crucial role in EDA by summarizing and providing meaningful insights into the dataset's main characteristics. They describe the central tendency, variability, and distribution of numerical variables. By analyzing descriptive statistics, researchers can understand the range, spread, and overall structure of the dataset, aiding in decision-making processes.

The project's ultimate objective is to deploy an accurate, efficient, and scalable recommendation system for Disney+ shows and movies. The system will utilize personalized suggestions based on user preferences and behavior, enhancing user engagement and potentially boosting the streaming platform's revenue. Performance evaluation using metrics such as precision, recall, and F1-score will ensure the system's effectiveness and allow for continuous improvement. By combining data science techniques, EDA, and the application of recommendation systems in the streaming media industry, this project seeks to contribute to the success and growth of Disney+ as a prominent player in the industry.

# References & Appendixes – Slide Presentation, Data, LR etc.

Ahuja, R., Solanki, A., & Nayyar, A. (2019). Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 263–268. https://doi:10.1109/CONFLUENCE.2019.8776969

Awan, M. J., Khan, R. A., Nobanee, H., Yasin, A., Anwar, S. M., Naseem, U., & Singh, V. P. (2021). A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. Electronics, 10(10), 1215. <https://doi.org/10.3390/electronics10101215>

Fortune Business Insights (2022) *Video Streaming Market Size Forecast 2022-2029*, *Video Streaming Market Size, Share, Growth & Forecast [2029]*. Available at: https://www.fortunebusinessinsights.com/video-streaming-market-103057 (Accessed: 08 May 2023).

Gomez-Uribe, C. A., & Hunt, N. (2015). The Netflix Recommender System. *ACM Transactions on Management Information Systems*, *6*(4), 1–19. https://doi.org/10.1145/2843948

Richter, F. (2021) *Infographic: Where Americans get their stream on*, *Statista Infographics*. Available at: https://www.statista.com/chart/25382/most-used-video-streaming-platforms/ (Accessed: 08 May 2023).

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. Journal of Big Data, 9(1), 59. https://doi:10.1186/s40537-022-00592-5

Schwartz, B. (2016) *The paradox of choice why more is less*. New York: Ecco.

Zhang, J., Wang, Y., Yuan, Z., & Jin, Q. (2020). Personalized real-time movie recommendation system: Practical prototype and evaluation. Tsinghua Science and Technology, 25(2), 180-191. https://doi:10.26599/TST.2018.9010118