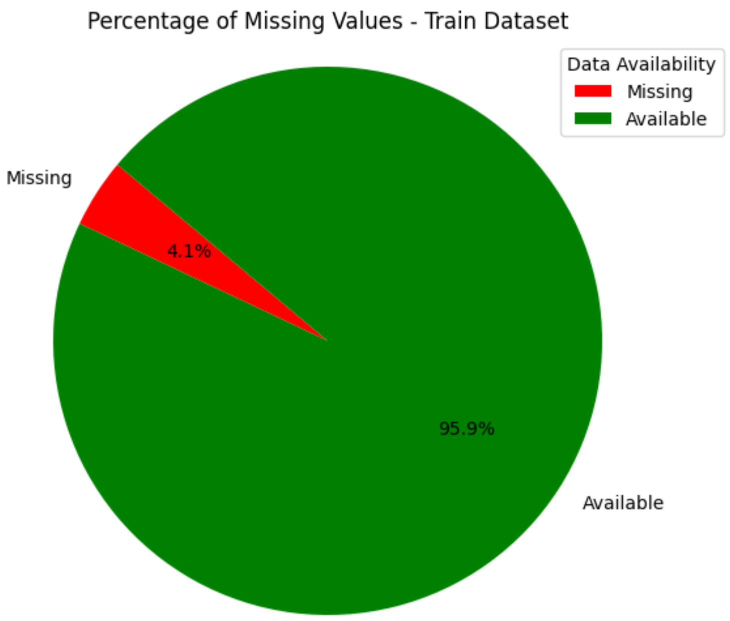
KOVAI.CO
ASSESSMENT
TASK - 1

By MAHALAKAME RM

1	Percentage of Missing values - Train Dataset
2	Distribution of show types
3	Distribution of movie release years
4	Distribution of TV Show release years
5	Content Additions by month
6	Top 10 Oldest and newest movies and TV Shows
7	Top Countries with the most content added
8	Distribution of content rating
9	Rating distribution for movies
10	Rating distribution for TV Shows
11	Distribution of Content Duration
12	Top Cast members by Number of content items
13	Top Directors of movie by Number of content items
14	Top Directors of TV Shows by number of content items
15	Top content Genres
16	Word cloud of common words in description
17	Content Added over time
18	Correlation Heat Map
19	Top 10 Movies by Duration
20	Top 10 TV Shows by duration

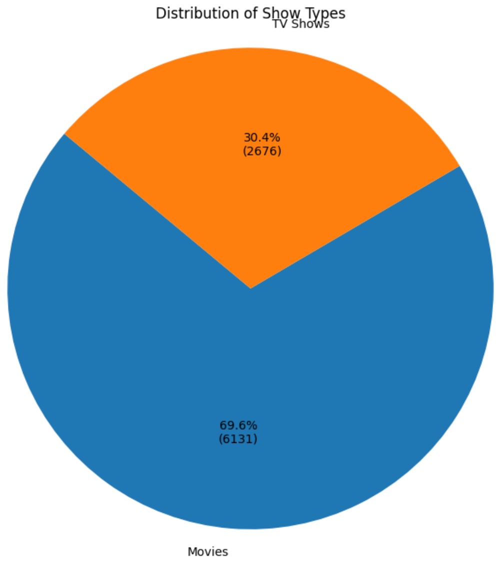
1. Percentage of Missing values - Train Dataset



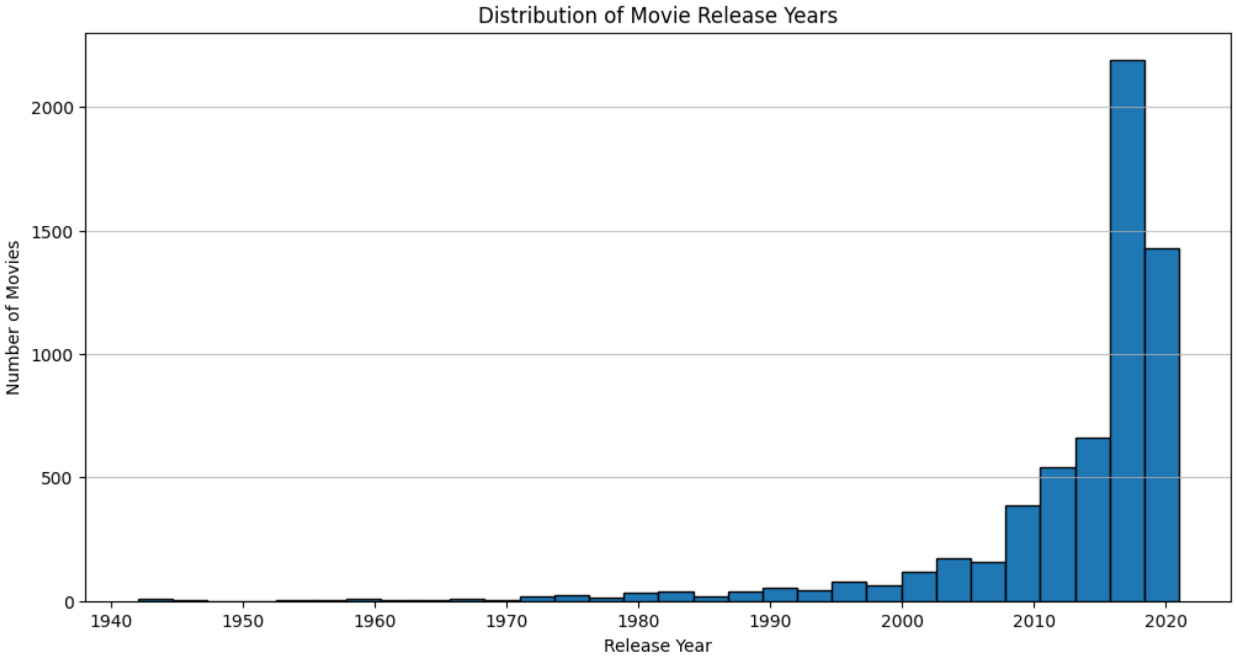
- Observation 1: The "director," "cast," and "country" columns have a high percentage of missing data, with 2634, 825, and 831 missing values, respectively, indicating the need for data imputation or considering how to handle these missing entries.
- Observation 2: "Rating" has 4 missing values, suggesting that some content may not have received a rating, which should be taken into account when conducting analysis involving content ratings.
- Observation 3: Only a small number of missing values (10) are present in the "date_added" column, which indicates that most entries have information about when they were added to the dataset.
- Observation 4: The other columns, such as "show_id," "type," "title," "release_year," "duration," "listed_in," and "description," appear to have no missing data, which is important for maintaining the completeness of these key attributes in your dataset.
- Observation 5: Approximately 41% of the data in the dataset is missing, highlighting the importance of addressing missing values and carefully considering how to handle them during the analysis to ensure the integrity of the insights derived from the dataset.

2. Distribution of show types

- The distribution of show types (as per the "type" column) in the dataset reveals that it predominantly contains a mix of TV shows and movies, which is a key characteristic to consider when exploring the content in the dataset.
- In this dataset, movie data represents approximately 69.9% of the content, while TV shows make up about 30%, indicating that movies are the predominant content type available, which is essential information for understanding the content distribution.

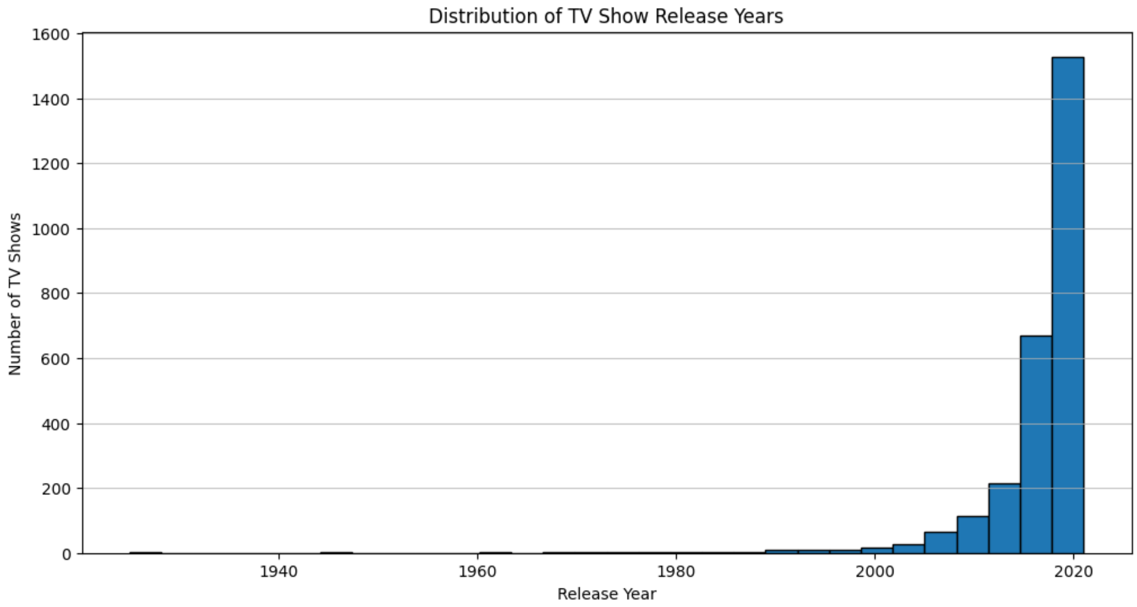


3. Distribution of Movie release years

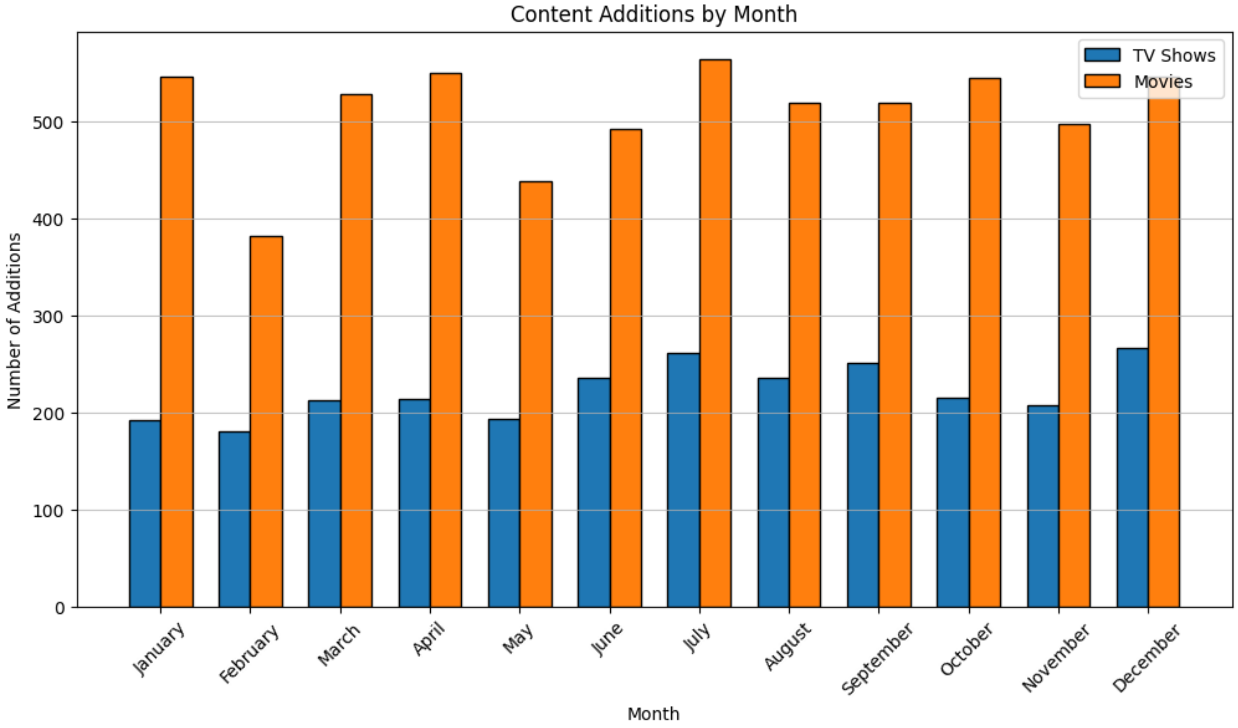


- The distribution of movies in the dataset shows a notable concentration of releases during the decade from 2010 to 2020, with 2021 exhibiting a significant increase in content. This trend suggests that the majority of movies in the dataset were released during this timeframe, with 2021 standing out as a year with a particularly high volume of content.

4. Distribution of TV Show release years

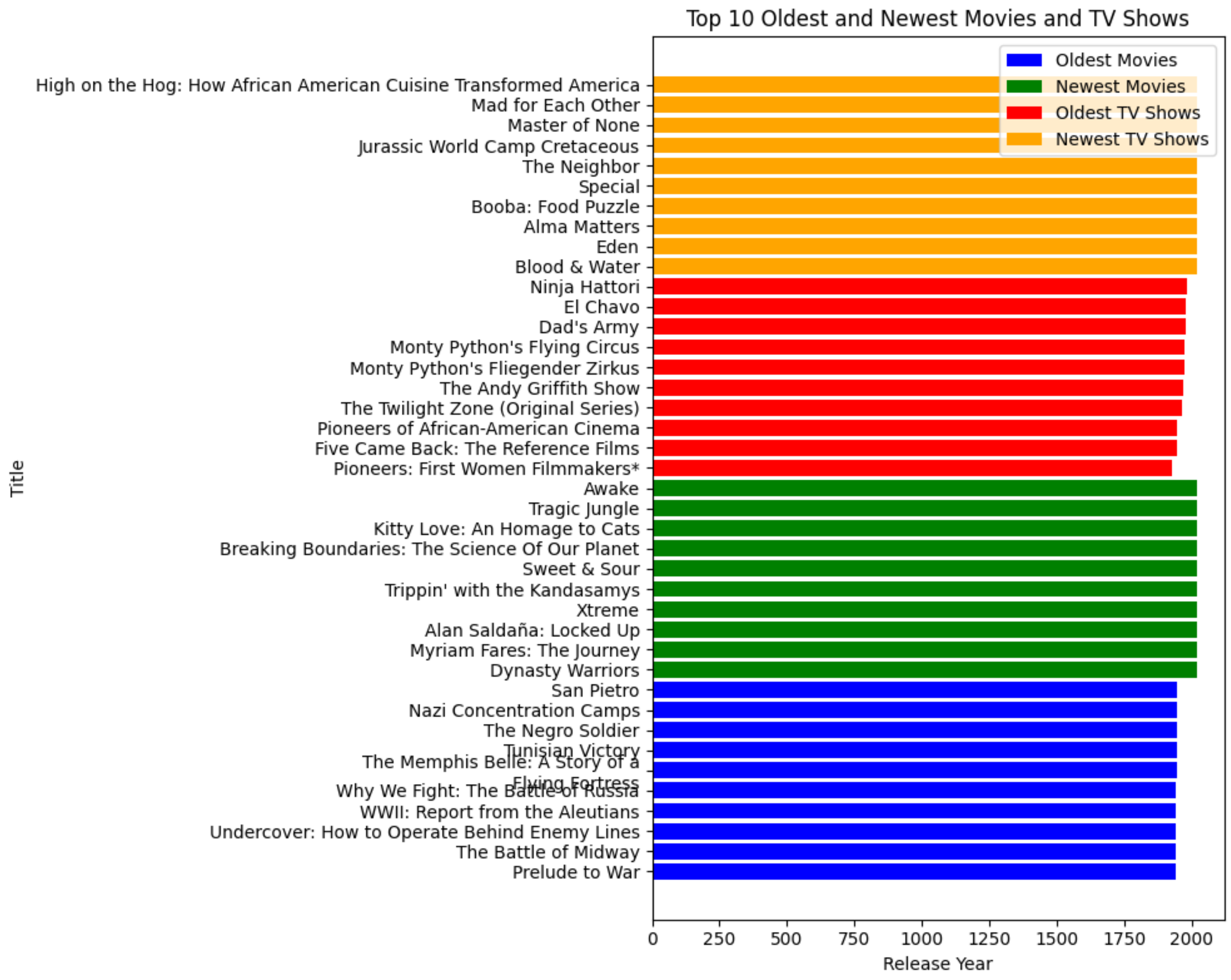


5. Content Additions by month



- The data reveals a consistent pattern in content additions by month, with a recurring and structured trend. This pattern may indicate that the dataset has a specific schedule or strategy for adding new content each month, which can be valuable for understanding the platform's content management approach.

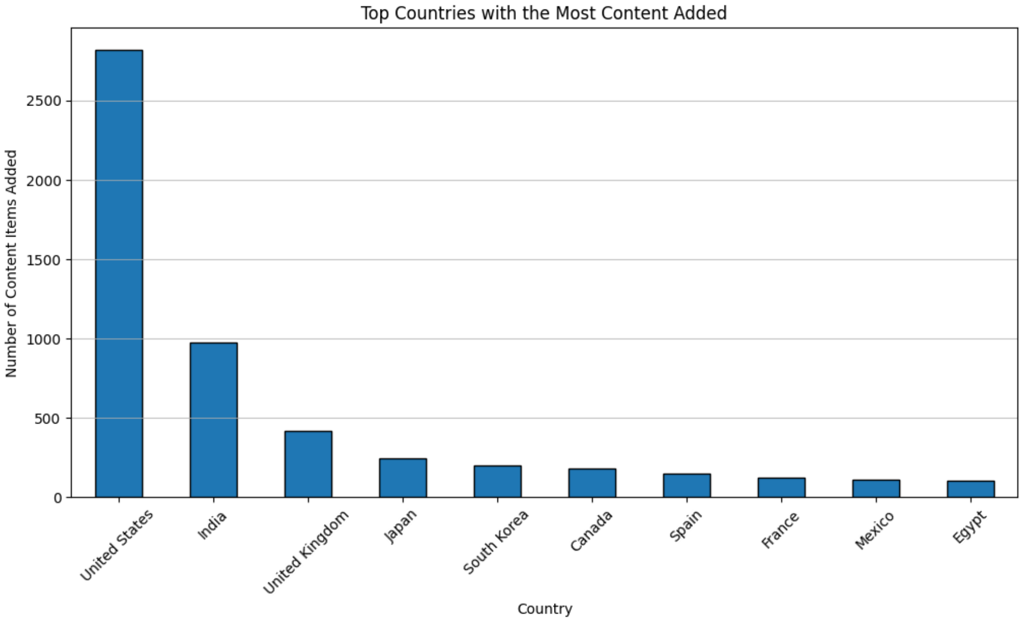
6. Top 10 Oldest and newest movies and TV Shows



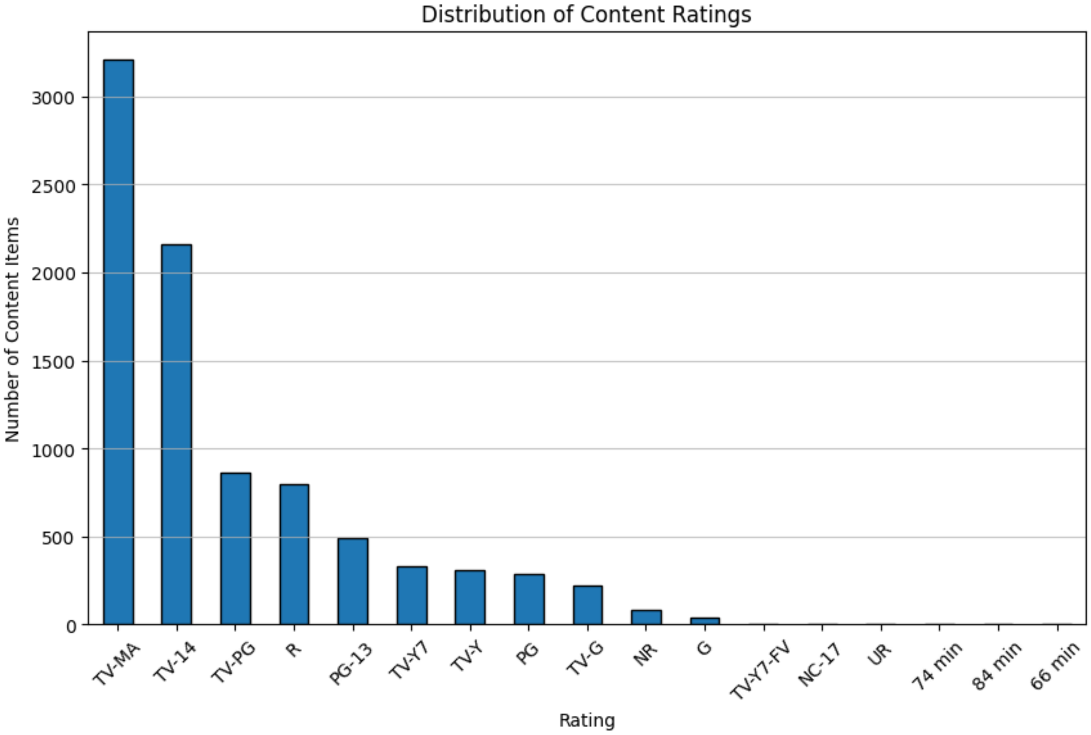
- The dataset provides information on the top 10 oldest and newest movies and TV shows. The oldest entries, dating back to the early 1940s and even 1925, showcase classic content. On the other hand, the newest entries, all from 2021, reflect the most recent additions, highlighting a wide range of content from different eras for viewers to explore.

7. Top Countries with the most content added

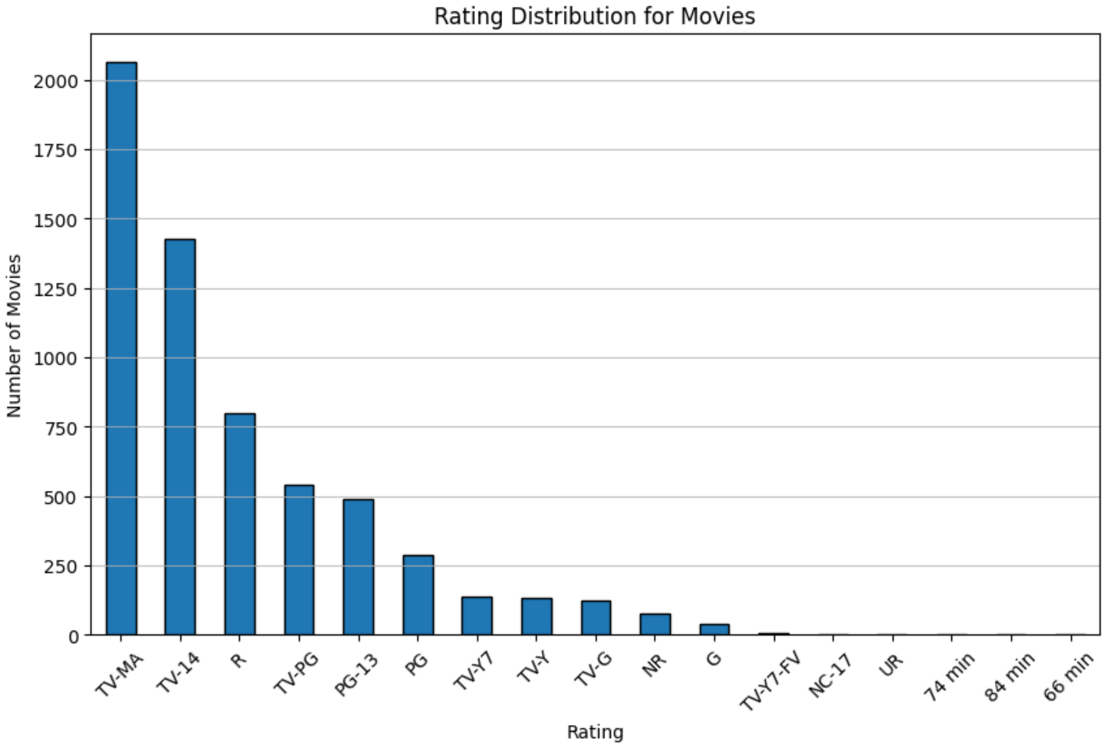
- The top three countries with the most content added to the dataset are the USA, India, and the UK. These countries have a significant presence in the platform's content library, suggesting a diverse range of content from these regions.



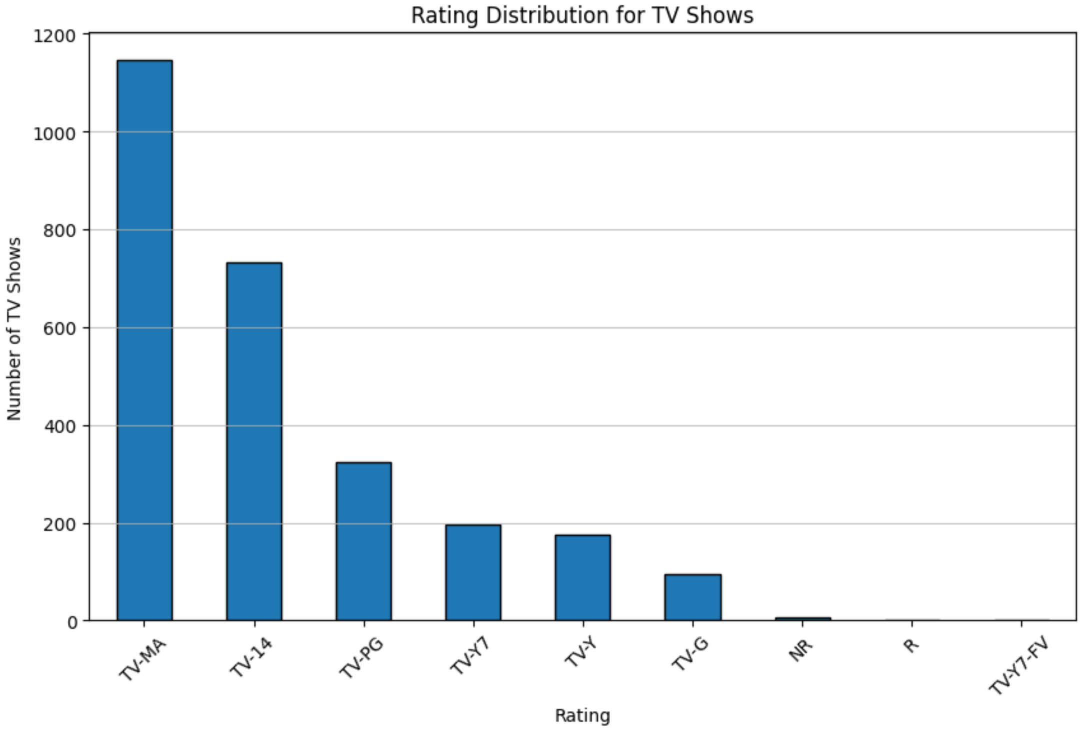
8. Rating distribution for TV Shows



9. Rating distribution for movies

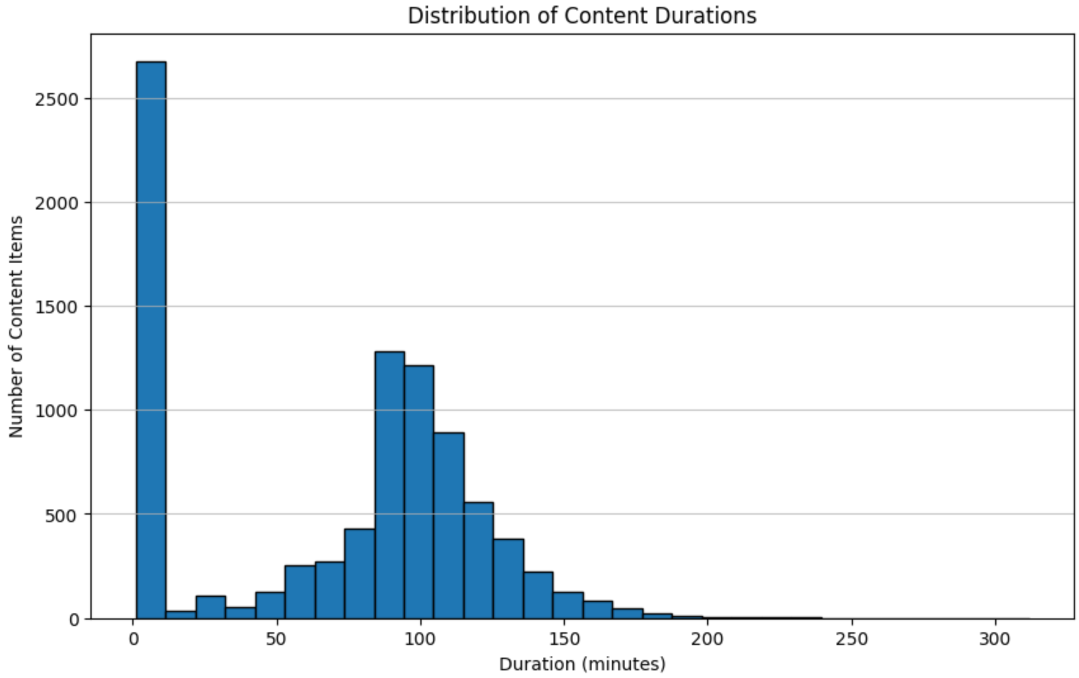


10. Rating distribution for TV Shows



- If we look at the ratings for movies and TV shows in the dataset, the ones with "TV-MA," "TV-14," and "TV-PG" are the most common. So, if you see these ratings, it tells you what kind of content to expect – whether it's suitable for adults, teenagers, or a general audience.

11. Rating distribution for TV Shows



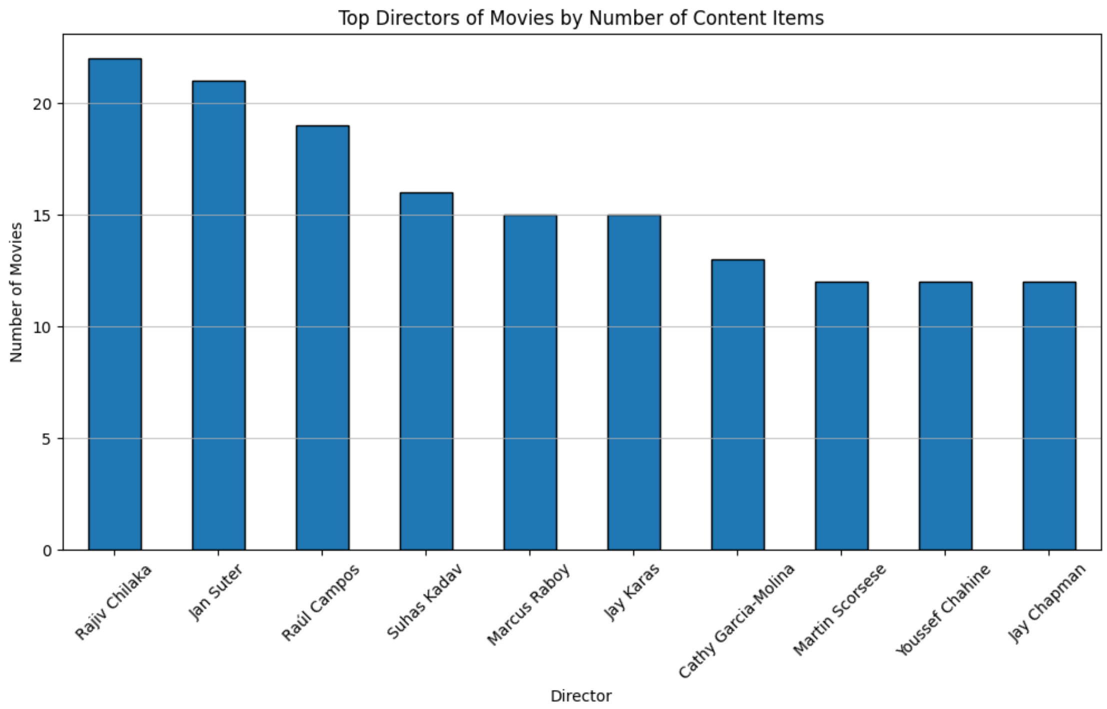
- Looking at the durations of movies and TV shows, many of them are between 0 to 10 hours, indicating that a significant portion of the content is relatively short and doesn't require a long time commitment.

12. Top Cast members by Number of content items

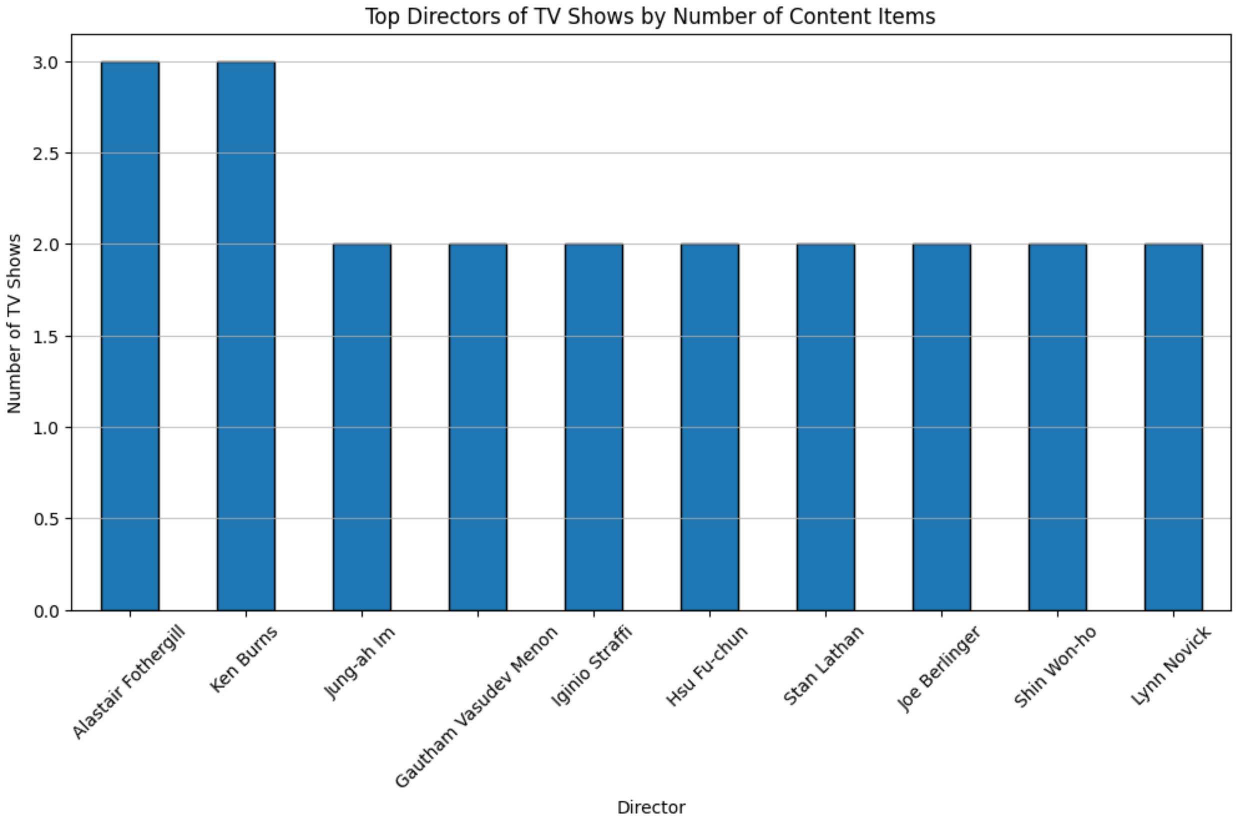
- The dataset allows us to identify the top cast members who appear in the most content items. This information can help us understand which actors or actresses have been prominently featured in a significant number of movies and TV shows.

13. Top Directors of movie by Number of content items

We can use the dataset to determine the top directors of movies based on the number of content items they've directed. This information provides insights into which directors have been involved in a significant number of movies in the dataset.

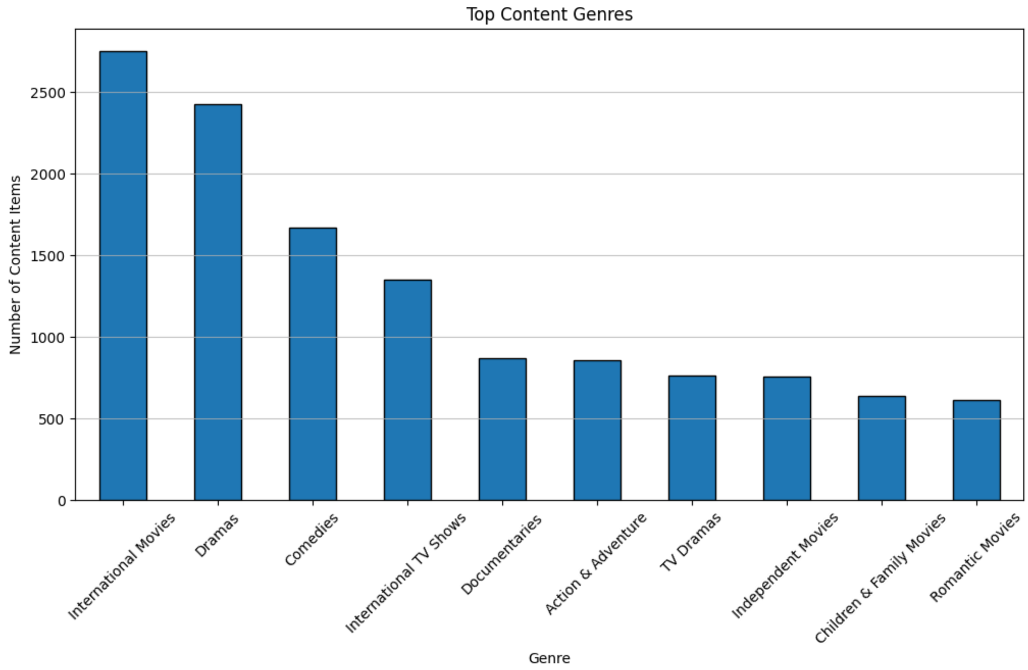


14. Top Directors of TV Shows by number of content items

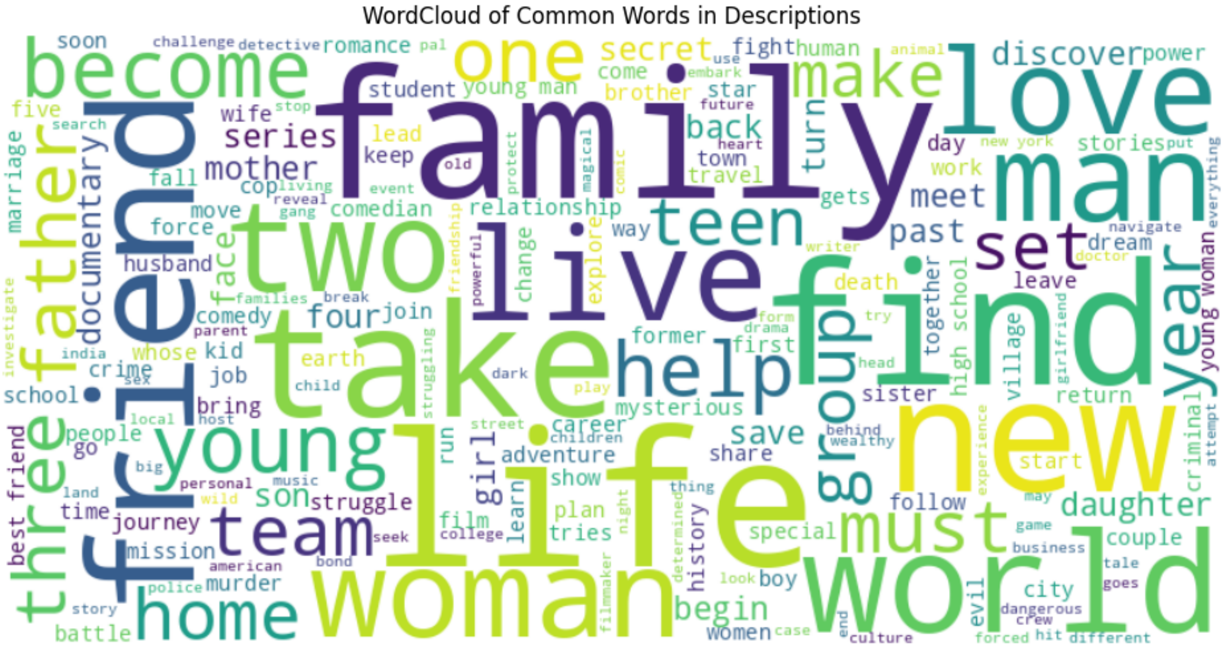


- We can use the dataset to determine the top directors of movies based on the number of content items they've directed. This information provides insights into which directors have been involved in a significant number of movies in the dataset.

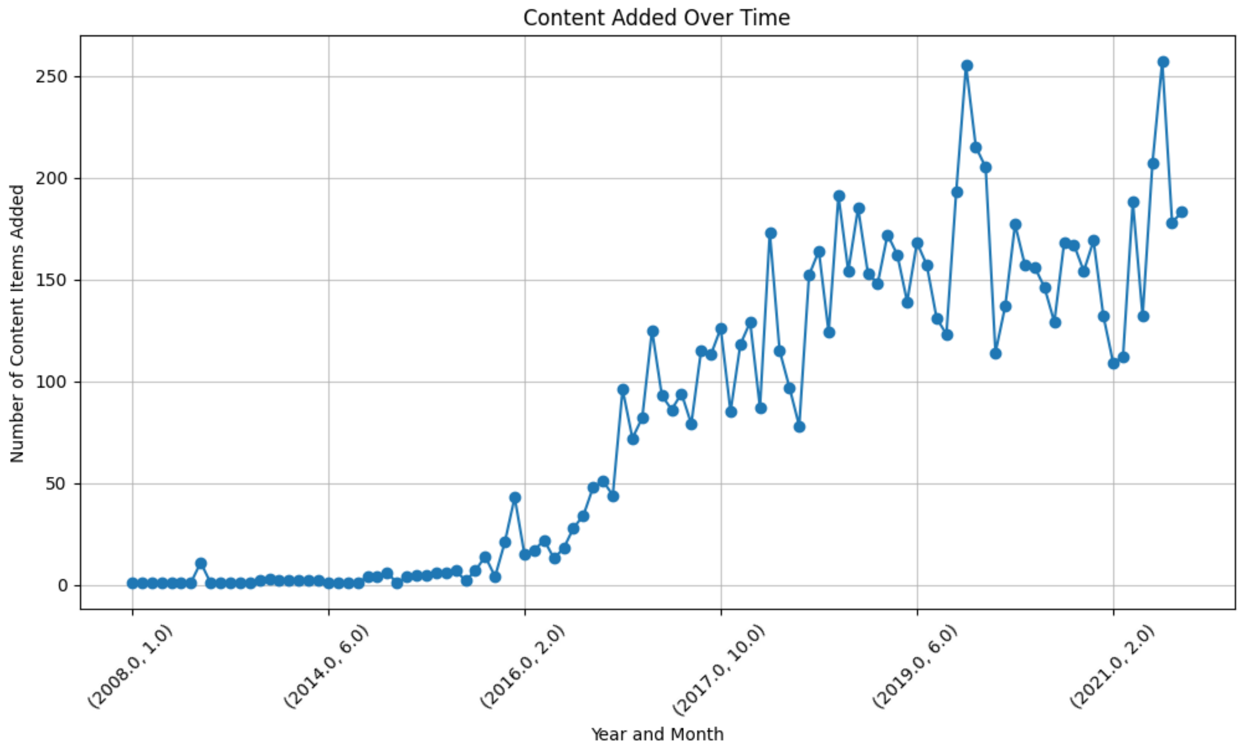
15. Top content Genres



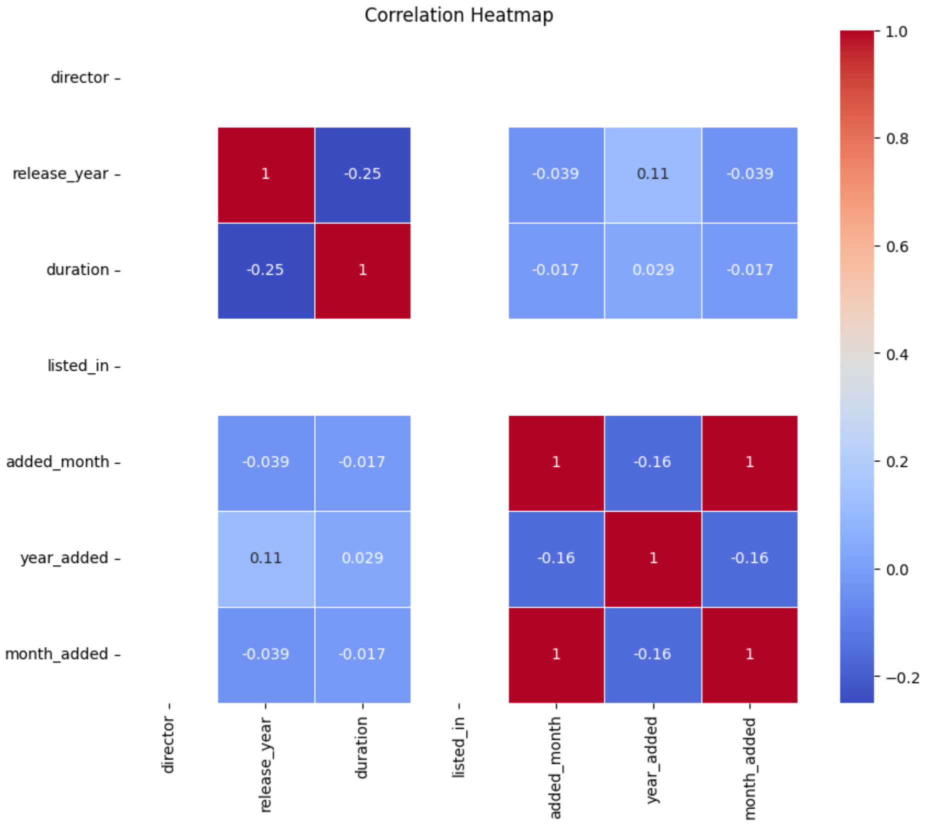
16. WordCloud of Common Words in Descriptions



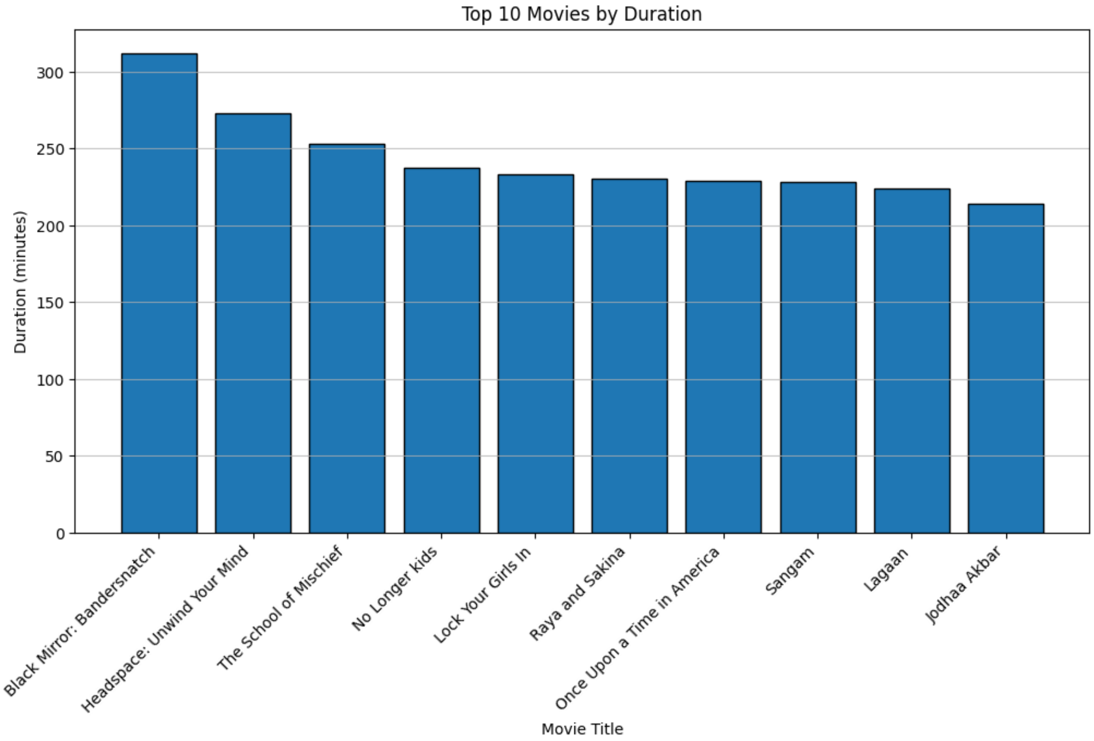
17. Content Added over time



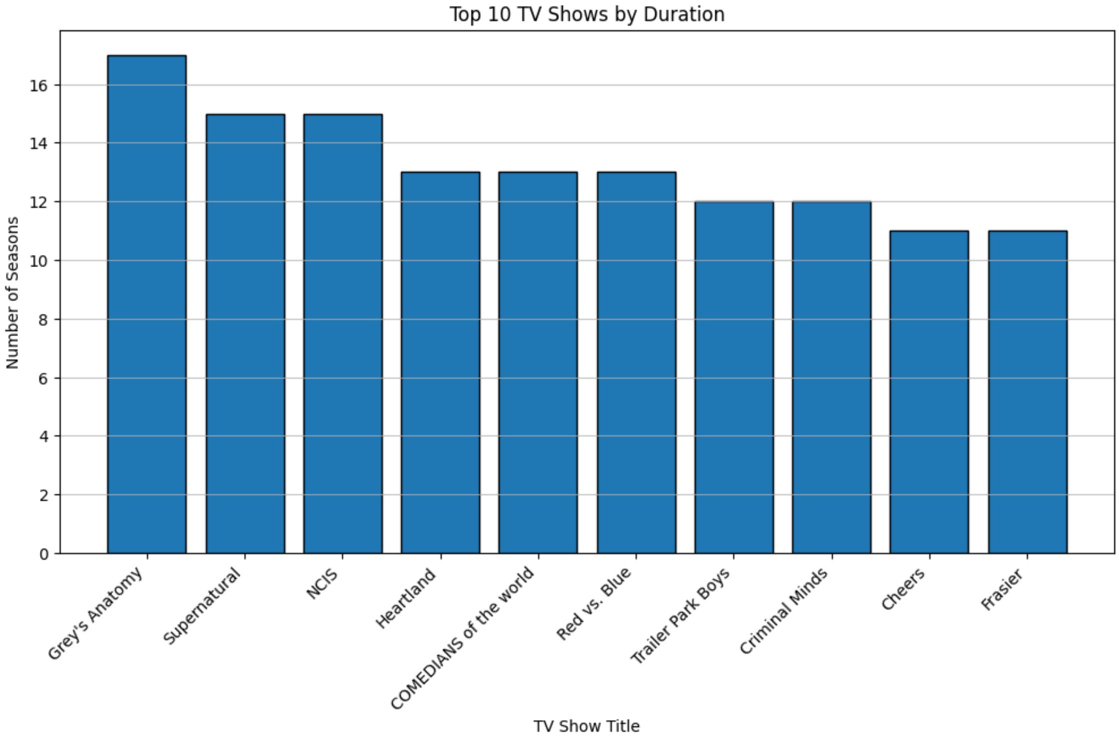
18. Correlation Heat Map



19. Top 10 Movies by Duration



20. Top 10 TV Shows by duration


Top 10 TV Shows by Duration
Number of Seasons
Grey's Anatomy
Supernatural
NCIS
Heartland
COMEDIANS of the world
Red vs. Blue
Trailer Park Boys
Criminal Minds
Cheers
Frasier
TV Show Title