**Cleaning step: Treating NULL in Overview, Keywords and Homepage.**

There were 3 NULL I values in the overview, 412 NULL values in Keywords, and 3091 in the homepage. Since the overview had only 3 NULL values in the overview, we treated NULL values by removing the rows that had NULL values in the overview. But we couldn't do the same because of this column's high concentration of NULL values. Since the homepage provides links to the movie's homepage, we decided to remove that feature instead of removing the rows. Keywords are the words used to search for the movie, and since they have many NULL values, we decided to remove the column.

**Cleaning step: Revenue and Released status (Post-Production and Rumored)**

The feature Revenue mentions the money made by the movie. Three unique status labels are Rumored, Post-Production, and Released. In this step, we checked if any of the movies that are labeled movies have a revenue of more than 0. It wouldn’t be correct if the movie has listed revenue when they are not released. Rumored movies didn’t have any instance of it. But post-production had one movie that had revenue more than 0 listed. So, we treated it by checking if the movie had been released by searching online, and it was, so we changed the status to Released.

**Cleaning Step: Revenue of Released Movies**

In this step, we checked if the Revenue of the released movies was more than 0 and since there were movies with 0 revenue listed and since Revenue listed the amount of money earned during the showings and streaming. We treated this by removing the rows with Revenue listed as 0 in the released movies.

**Cleaning Step: Original title and Translated Title**

The movies in the dataset are from different countries, so the original title is in a different language. We decided to use the Translated title instead of the original to keep everything in the same language.

**Cleaning Step: Most Occurring word in Keywords.**

To find the most occurring words in keywords, we had to split the text of each row, remove stopwords, and format in a string to get results from the word cloud plot.

**EDA: Overview sentiment Analysis**

The overview of the movies is supposed to be neutral and not opinionated. We did sentiment analysis on overview to check the sentiment and how opinionated. It is understood that the mean is 0.042329 with a standard deviation of 0.21. This tells us a good amount of overview is almost neutral, and more overviews have a positive than negative sentiment. We also did a Subjectivity analysis and found that overviews are more opinionated than factual or neither, as the mean is 0.46764 and the standard deviation is 0.22.

A blue graph with white text

Description automatically generatedA blue graph with white text

Description automatically generated

**EDA: Popularity**

We first used describe() to get the mean, std, min, values in each quartile, and max values. Mean was 28.608220, and standard deviation was 35.60820, suggesting that the data is more spread out. We plotted a histogram to check the distribution of popularity. The Y-axis is log-scaled, and many movies are not very popular. We also checked the average popularity of movies in each month of the year, and from the graph, the popularity of movies is higher during the months of June, May, and July, which are the summer months. A scatter plot was plotted to see how revenue is related to popularity. We can see that the popularity is higher when there is higher revenue. Although there are outliers that have smaller revenue, they are more popular, and vice versa.

A graph with blue squares

Description automatically generated

A graph with blue lines

Description automatically generated

A graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated

**EDA: Status**

Status has three unique values: Released, Rumored, and Post-production. Released movies have the highest frequency in the data set. We plotted a scatter plot for each category between vote count and popularity. We can see that in Popularity increases as the vote count increases, but for Rumore and Post-Production, that is not the case popularity is not related to the Vote count.

A graph of a scatter plot

Description automatically generated with medium confidence

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

**EDA: Vote Average**

Vote Average is the average ratings of the user, and vote count represents the no. of people who voted. The movies have an average rating of 6.31, as seen when running describe(), with a standard variation of 0.892. We plotted a histogram to see the distribution, and it can be seen the majority of data is concentrated near the average of 6.3. From the bar plot that calculates the vote average per month, it can be seen the movies released during the months of December and May have a higher vote average than other months.

A graph of a distribution of vote

Description automatically generated

A graph of a distribution of vote

Description automatically generatedA graph with blue lines

Description automatically generated

**EDA: Trend of revenue made by Movie in each Month**

As part of EDA we created a line plot to look at the trend of what the average revenue for the movie in each month. We can see the average revenue is the lowest in September and January month in our dataset.

A graph with blue lines

Description automatically generated

**EDA: Most occurring terms in Keywords**

Keywords are words used to describe a movie and what the movie is related to. The word cloud shows the most occurring terms in the Keywords, which tells us the frequently occurring theme or characteristic of the movie.

A close-up of words

Description automatically generated