Movie Market Prediction

Problem Statement : Estimating the revenues and popularity of a film.

a) The global film industry is dynamic and intensely competitive. Success hinges on various factors, from creative elements to strategic marketing and business decisions. Predicting a film's revenue and popularity becomes crucial as it empowers stakeholders—such as production companies, distribution firms, and investors—to navigate this complex landscape with better foresight and precision. It helps them make informed decisions, allocate resources effectively, and stay ahead in an ever-evolving industry. Predicting revenue and popularity is significant because it directly impacts financial outcomes, enables more strategic decision making, reduces the risks of getting losses, and ensures the efficient allocation of resources.

b) The film industry involves substantial financial investments. Production costs, marketing expenses, and distribution budgets are considerable. Accurate revenue prediction is crucial to ensure that these investments yield profitable returns. The scope of this project is to help the production companies, investors across the globe in identifying the target audiences and optimizing advertising campaigns and also to allocate budgets more effectively, reduce the financial risks, and also optimize the allocation of different resources. This can leverage the revenue prediction to enhance their financial decision making and enhance the viewers experience too.

Data Source:

The Dataset we've used is a movie dataset present over the web for free. It contains data of over 4,800 movies. It contains a total of 24 features. The list of features and their types are as follows:

a) Index: numerical(integer)
b) Budget: numerical(integer)
c) Genres: text
d) Homepage: text(url)
e) Id: numerical(integer)
f) Keywords: text
g) Original_language: categorical(text)
h) Original_title: text
i) Overview: text
j) Popularity: numerical(float)
k) production_companies : text(json)

l) Production_countries: text(json)
m) Release_date: temporal data(year-month-day)
n) Revenue: numerical(integer)
o) Runtime: integer(minutes)
p) Spoken_languages: text(json)
q) Status: categorical(text)
r) Tagline: text
s) Title: text
t) Vote_average: numerical(float)
u) Vote_count: numerical(integer)
v) Cast: text
w) Crew: text(json)
x) Director: text

In total eight columns in the dataset have null values in them and it is also rife with zero-value rows in the majority of the columns.

Data Cleaning:

In the process of preparing our dataset for analysis, we conducted comprehensive data cleaning and transformation tasks. The main objectives were to handle null and NA values, create new columns for analysis, standardize date formats, remove unnecessary symbols, and convert 'runtime' into a more readable format. The following data cleaning strategies were carried out in the dataset:

a) Treating Genre Column:

Examples of data in the original Genre Column.

0 Action Adventure Fantasy Science Fiction
1 Adventure Fantasy Action
2 Action Adventure Crime
3 Action Crime Drama Thriller
4 Action Adventure Science Fiction

A Single Movie is categorized into multiple genres; but they are kept in the same column. We found all the categories of genre present in the dataset by parsing the genre column; and made a separate column for each category where we filled the columns with binary encodings based on presence. We Found 22 different categories in total. *(TV, Movies)* and *(Science, Fiction)* are the same category but parsed as different words by our initial

pre-processing. So, we reworked them, and treated them as "TV Movie" and "Science Fiction" categories. Thus, altogether we had 20 different categories of genre, which we represented as different columns.

b) Treating Production_Companies Column

Examples of data in this column:

[{'name': 'Ingenious Film Partners', 'id': 289}, {'name': 'Twentieth Century Fox Film Corporation', 'id': 306}, {'name': 'Dune Entertainment', 'id': 444}, {'name': 'Lightstorm Entertainment', 'id': 574}]

[{'name': 'Walt Disney Pictures', 'id': 2}, {'name': 'Jerry Bruckheimer Films', 'id': 130}, {'name': 'Second Mate Productions', 'id': 19936}]

[{'name': 'Columbia Pictures', 'id': 5}, {'name': 'Danjaq', 'id': 10761}, {'name': 'B24', 'id': 69434}]

This Feature is a json that mentions all the production companies involved. These are text features and we converted it into a new feature that demonstrates no. of production companies involved in making the movie.

c) Filling Missing Values(Replacing Null) in 'runtime' column

Two records for the *runtime* column have missing values, which we filled by looking up imdb data.[1,2]

d) Filling Missing Values in(Replacing Zero) in Popularity column using mean popularity value

A few records for popularity have values of zero, which was replaced with mean popularity.

e) 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.

f) 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 instances 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.

g) Revenue of Released Movie

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.

h) 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.

i) Overview Cleaning

We Converted the text values of the overview column into quantitative measures of sentiment and subjectivity and created columns for each of them so that it could be used to explore other quantitative features like revenue, popularity.

j) Most Occurring words 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.

k) Treating Release Date Column

We Identified and treated null values in the release date columns. We also extracted Year and Month from the release date column, creating new columns for easy analysis. We converted the date format to a more standardized representation for consistency in the dataset.

l) Symbol Removal:

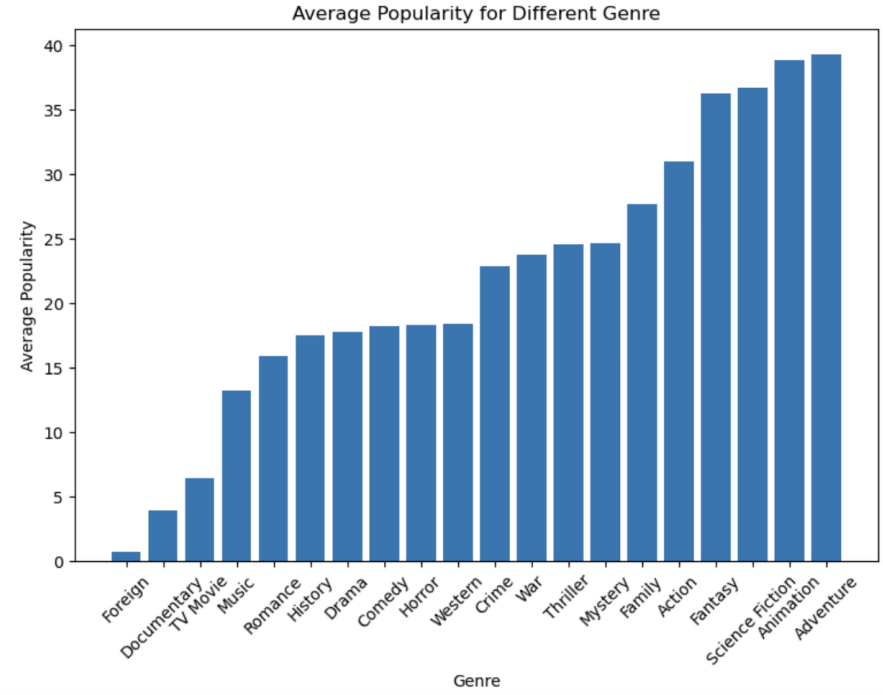
Removed unnecessary symbols, such as dollar signs, from relevant columns. Ensured uniformity for numerical columns, facilitating straightforward numerical analyses.

m) Treating 'Runtime' Column

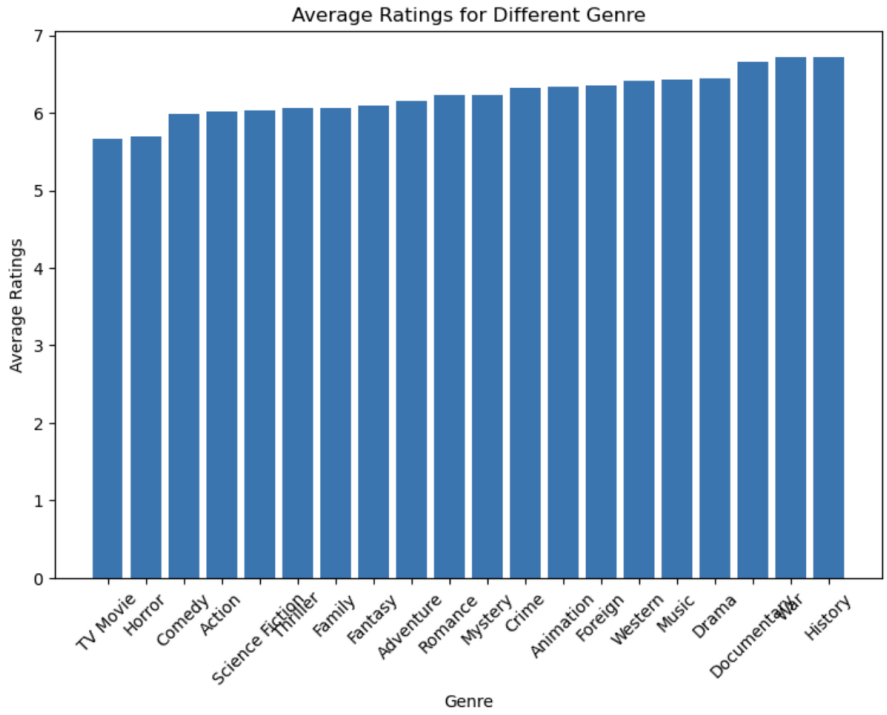
Standardized the runtime format into minutes for a more consistent and understandable measure. Ensured the runtime data is now in a readable format.

Exploratory Data Analysis(EDA):

a) Avg popularity for various genre

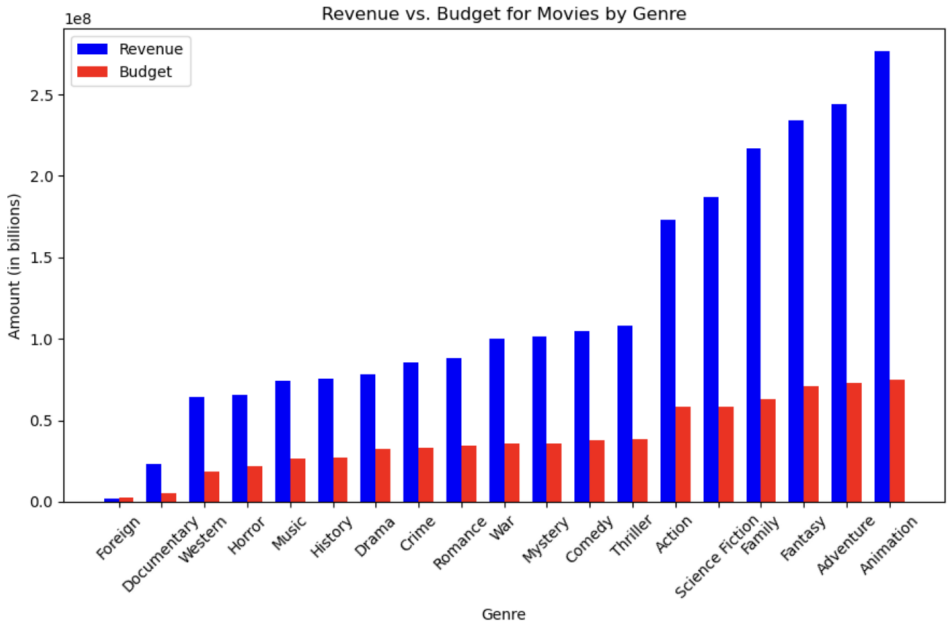


b) Average Ratings for various genre

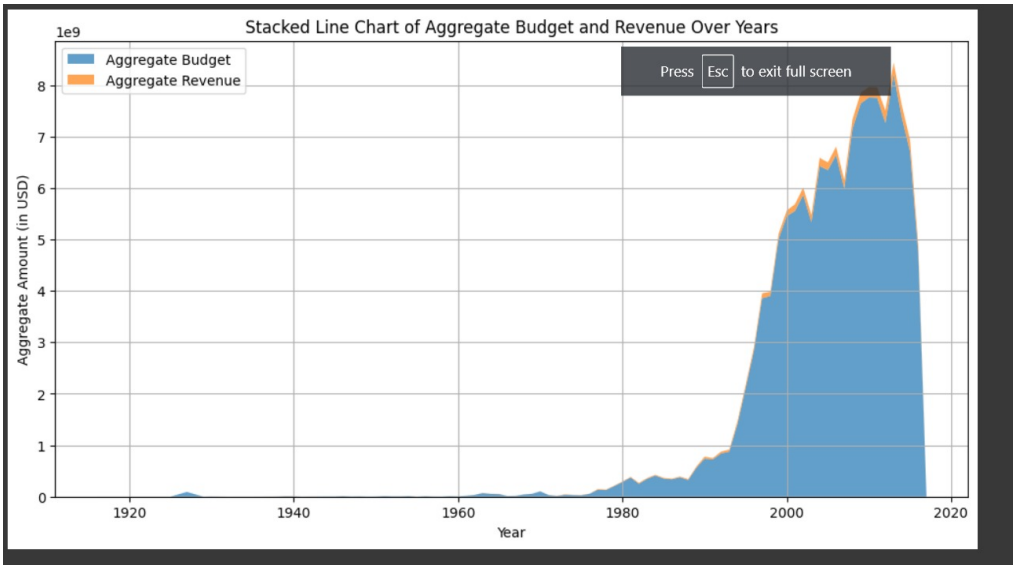


Movies that are highly popular generally have lower ratings, and critically acclaimed movies are less popular.

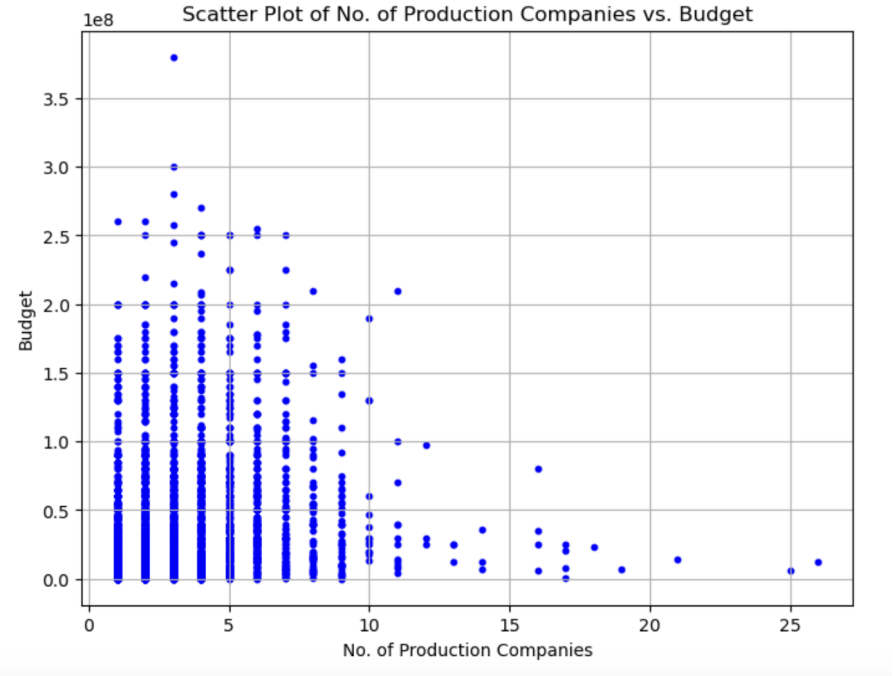
c) Average Revenue vs budget for movies by genre



Animation, Adventure, Fantasy, Science Fiction movies generally give better return on investment compared to other genres.

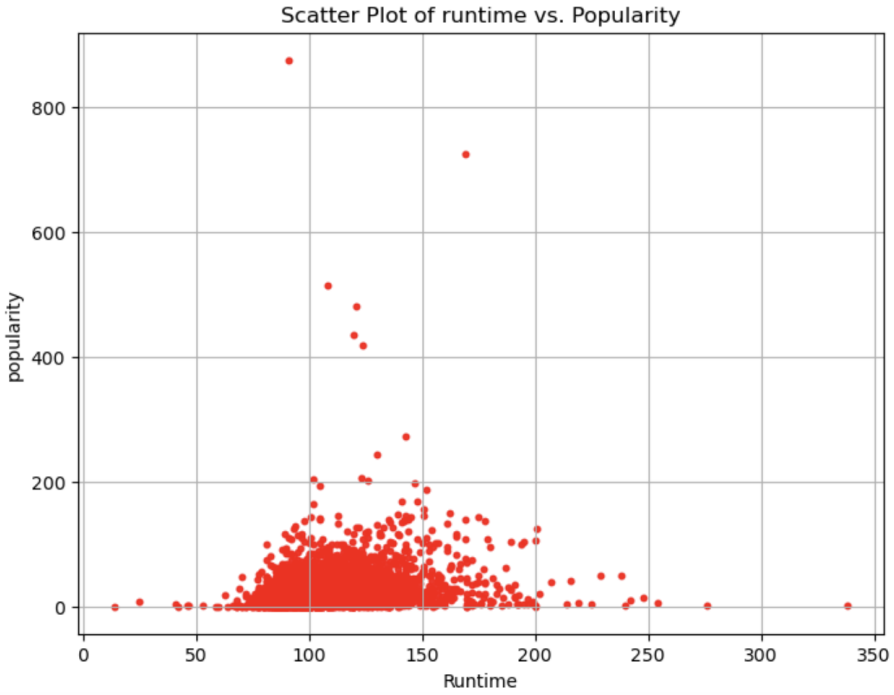


d) No. of Production Companies vs Budget



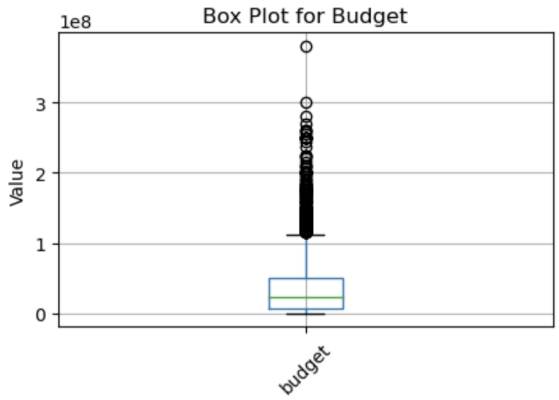
We found movies with over 20 production companies involved too. Large No. of production companies are seem to be involved because of difficulty in gathering production budgets.

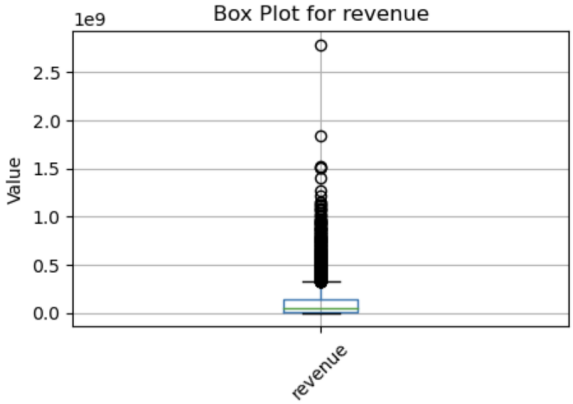
e) Runtime vs Popularity

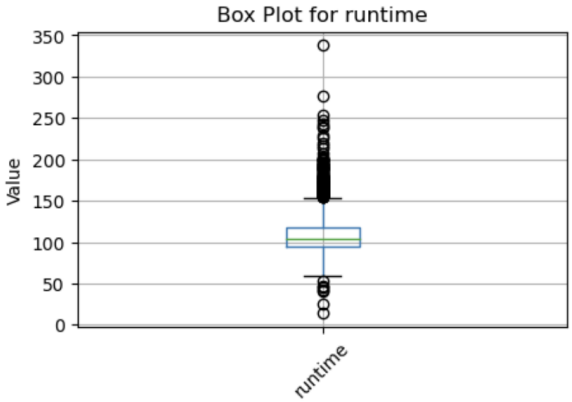


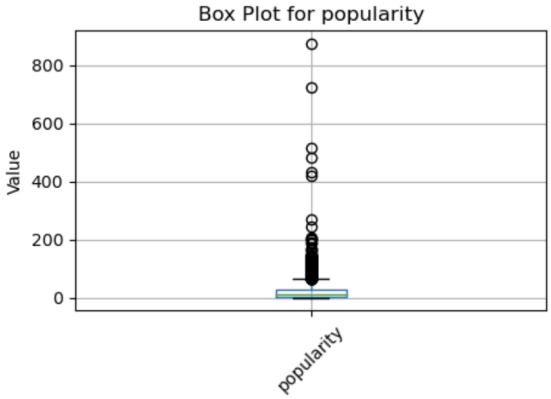
Movies that have lengths of around 1.5-2.5 hours are more popular.

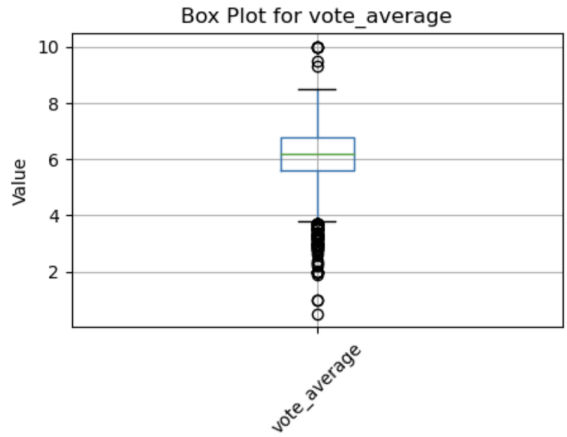
f) Visualizing the distribution of numerical features using Box plots







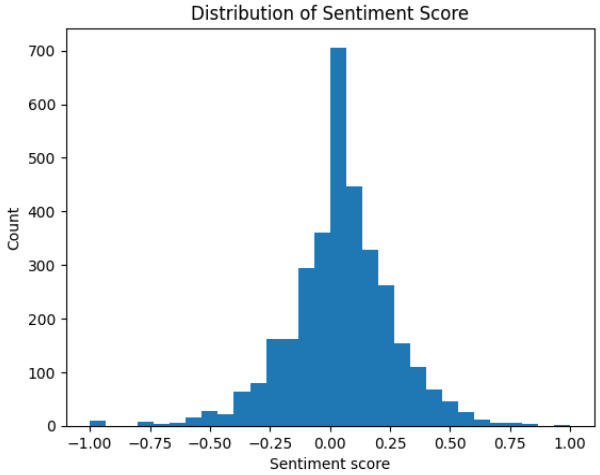


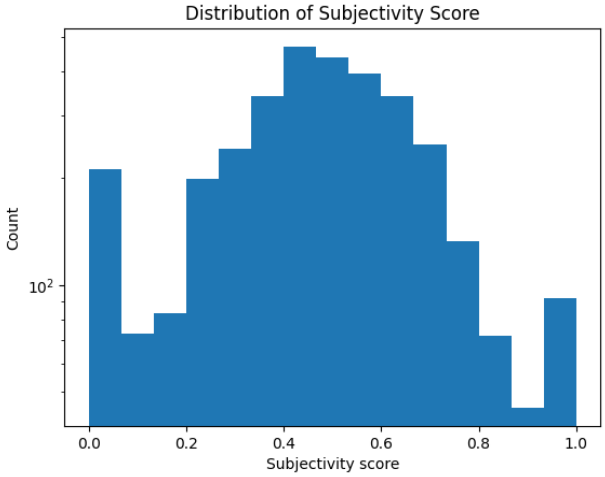


g) Overview sentiment Analysis

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. We found

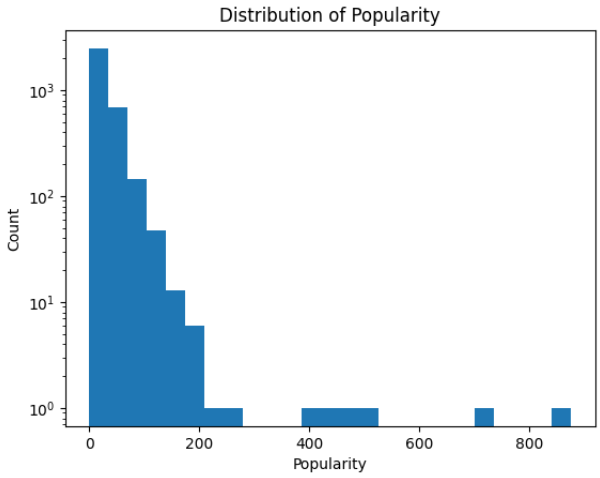
a mean of 0.042329 with a standard deviation of 0.21. This tells us a good amount of overview is almost neutral, and more overviews have a positive than negative sentiment. We also did a Subjectivity analysis and found that overviews are more opinionated than factual or neither, as the mean is 0.46764 and the standard deviation is 0.22.

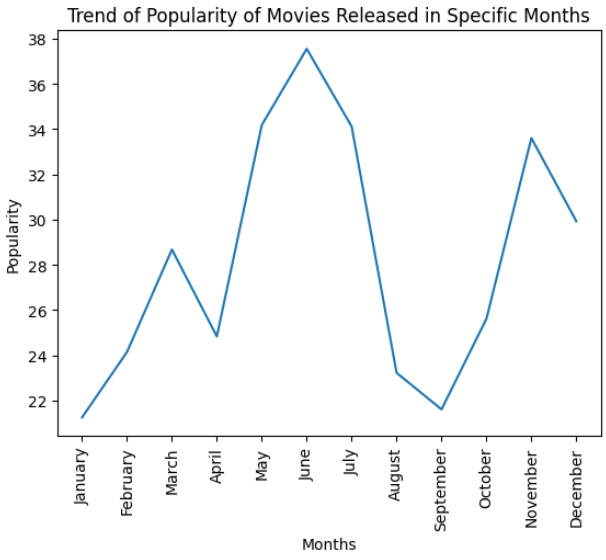


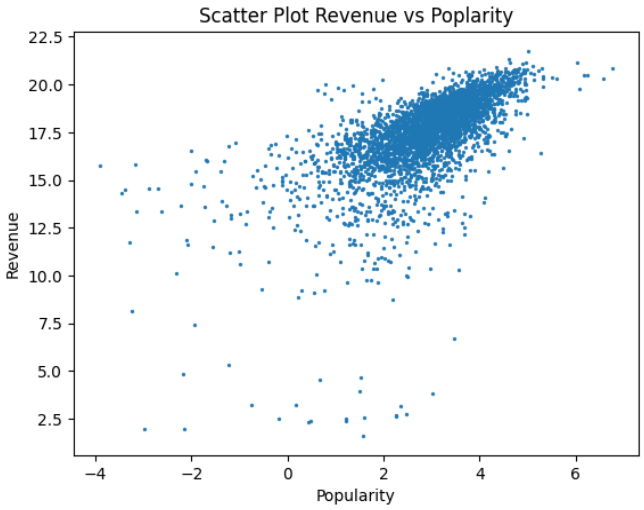


h) Popularity

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.

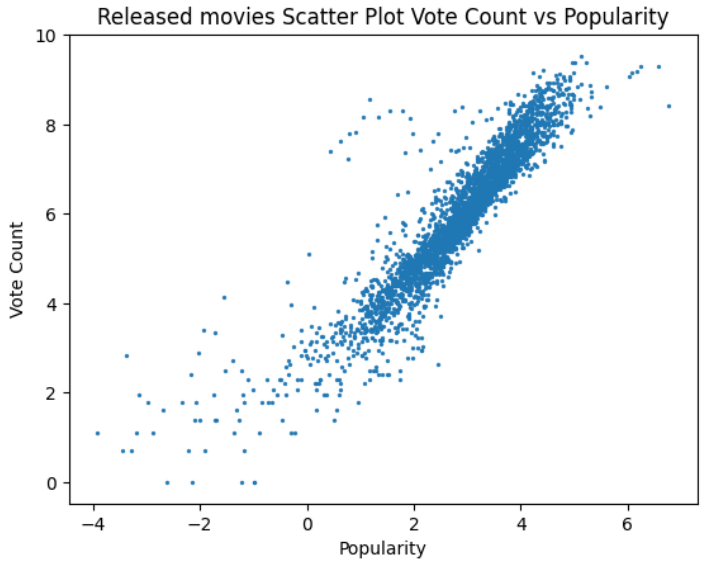






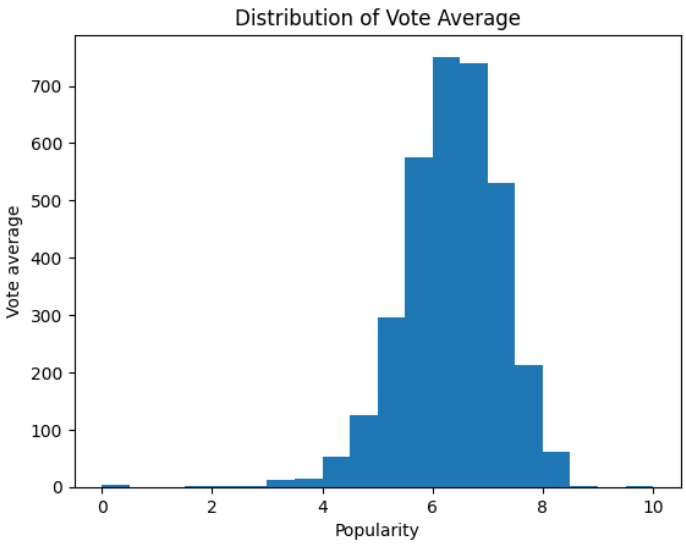
i) EDA: Movie Status

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 Popularity increases as the vote count increases, but for Rumored and Post-Production, that is not the case popularity is not related to the Vote count.



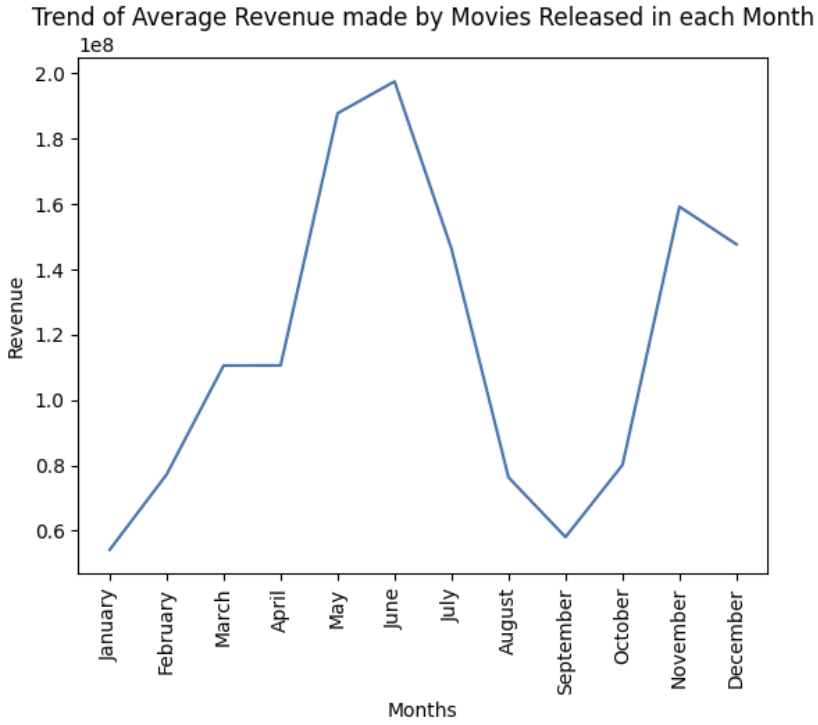
j) EDA: Vote Average

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.



k) Trend of revenue made by Movie in each Month

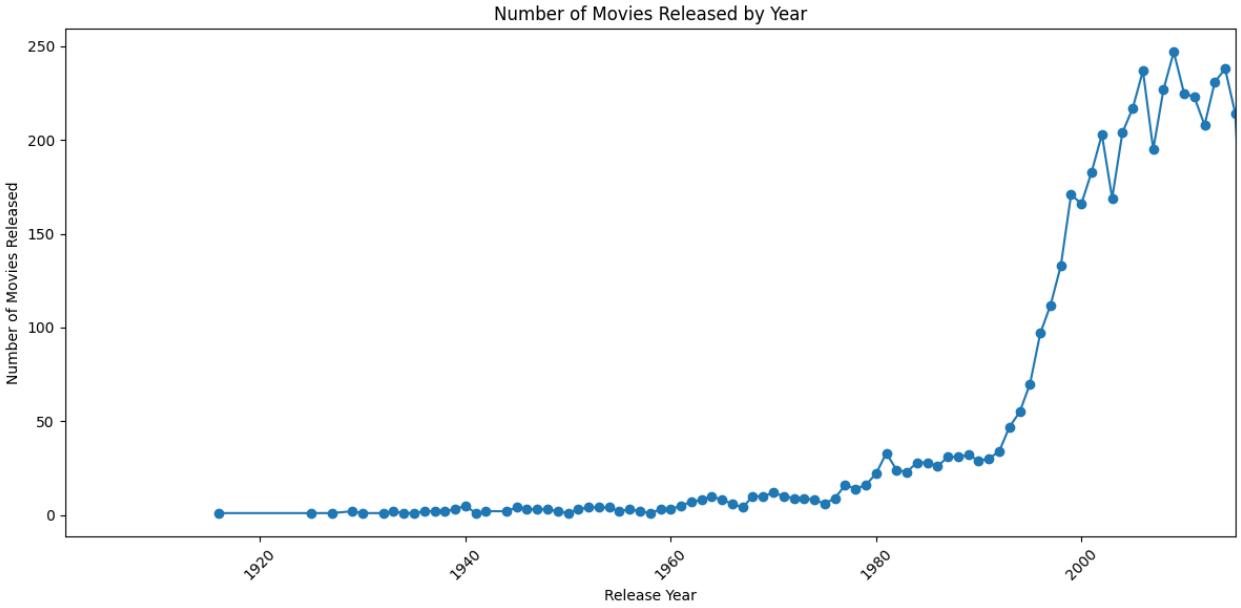
As part of EDA we created a line plot to look at the trend of the average revenue for the movie in each month. We can see the average revenue is the lowest in September and January month in our dataset.



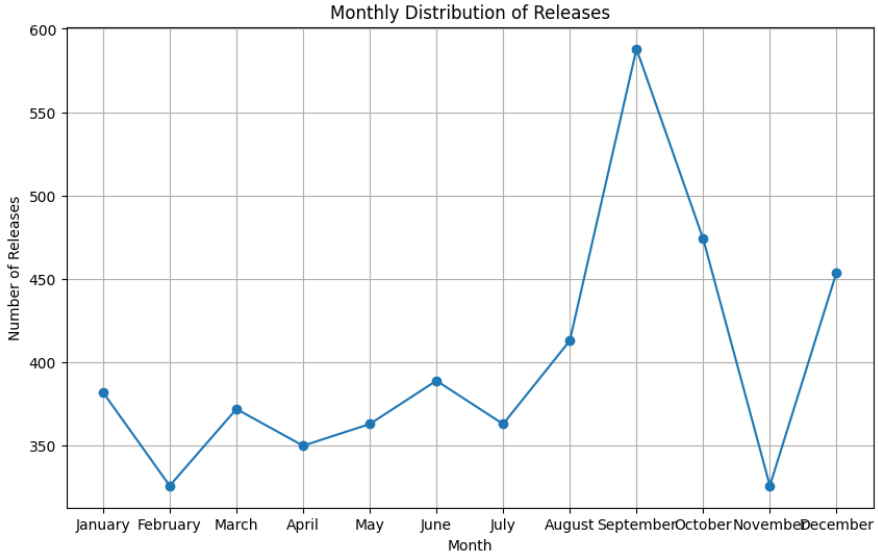
l) Most occurring terms in Keywords

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

m) Time series plot - Number of Movie releases - Year



n) Seasonality Analysis Monthly

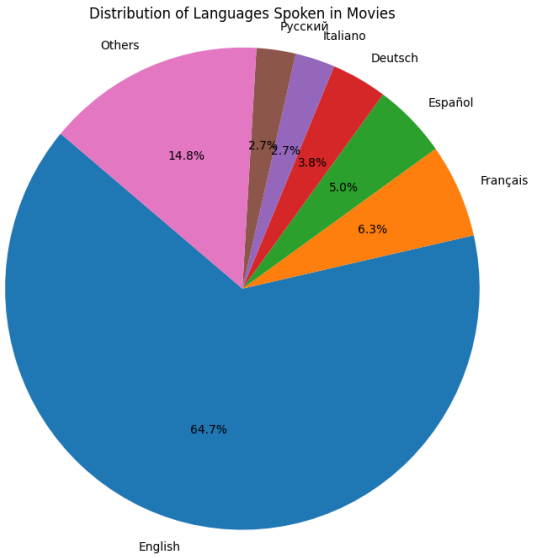


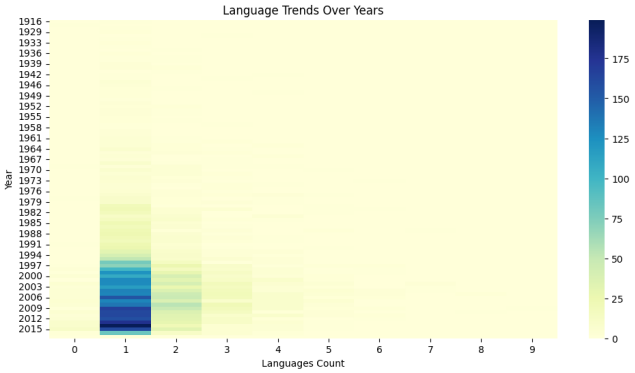
We saw that during september- December the majority of the movies release.

o) Weekend vs. Weekday Releases

(-3.139337589399419, 0.0017035818564200 19) - Based on the t-statistic and p-value, we can say that there is a significant difference in movie revenue between weekday and weekend releases, with weekend releases having a higher mean revenue.

p) Languages frequency - Movies





References

[1] [https://www.imdb.com/title/tt4704314/]

[2] [https://www.imdb.com/title/tt3856124/

[3] Dataset: https://raw.githubusercontent.com/rashida048/Datasets/master/movie_dataset.csv

[4] https://www.datacamp.com/tutorial/wordcloud-python

[5] NIST on EDA, https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

[6]https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524