DATA STORYTELLING PROJECT

A Data-Driven Journey into Movie Success Stories: IMDB Data Analysis

BY OLUWATOMISIN EBUN AROKODARE

The Internet Movie Database (IMDb) is a huge informational resource that offers details on a huge number of movies as well as their ratings, Meta score, gross, year of release, votes, movie duration, movie genre, movie directors and stars. The Internet Movie Database has become a major player in the world of film industry study and appreciation. IMDb's rating system, which allows viewers to assign a score to movies from 1 to 10, is one of the distinguishing characteristics of the website.

IMDb data analysis provides underlying trends and attributes that contribute to movie success stories, revealing audience preferences, critical reactions to movies and social trends that have shaped the film industry. In this project, I conducted a data-driven inquiry into the world of movie success stories using the capacity of IMDb data research.

Moving forward to the main aim of this project, I will be analyzing the top 250 movies in the internet movie database to gather insights.

Dataset Overview

The dataset that was used for this analysis was web scraped from the IMDB website; a popular online platform that was established in 1990. It serves as the most comprehensive movie database in the world, offering information about movies, TV episodes, and celebrities to professionals, movie fans worldwide, enthusiasts and researchers allowing them to explore their favorite shows and stay updated with the latest releases. The platform also keeps users informed about upcoming movies, trailers, industry news, and events. It offers user engagement features that allow users to make lists of their favorite movies, keep track of their viewing history, and join in discussion boards to share their views and ideas with other movie fans. The Top 250 selected IMDB ranging between 1978 to 2019 were analyzed.

Gathering (Scrapping) the data

The python request library with its methods was used to get the data from the IMDB URL(https://www.imdb.com/search/title/?count=250&groups=top_1000&sort=user_rating) and the content of the URL is parsed via Beautiful Soup. Matplotlib and seaborn libraries were imported for visualization purposes. For loop function was used to extract the content of the URL, and then returned it as a Pandas data frame using the panda's library.

Data Cleaning

1)From the top 250 movies scraped from the IMDB websites, movies genres appeared to have (3 genres for some movies while others have 2 genres). To have a consistent dataset, I split the movie genre to see only the first movie genre for each movie.

	0	1	2
0	Drama	None	None
1	Crime	Drama	None
2	Action	Crime	Drama
3	Biography	Drama	History
4	Crime	Drama	None
...	...	...	...
245	Drama	War	None
246	Animation	Action	Adventure
247	Action	Comedy	Crime
248	Drama	None	None
249	Action	Drama	Sport

250 rows × 3 columns

```
Out[15]: Action       61
         Drama        44
         Crime        37
         Biography    22
         Animation    20
         Comedy       20
         Drama        19
         Adventure    18
         Mystery       4
         Horror        3
         Western       1
         Horror        1
         Name: Movie_Genre, dtype: int64
```

2)The top 250 movie column names consist of trailing spaces, so I removed all trailing spaces. The Meta score column was converted from object to category data type, movie votes column was converted from object to Int datatype, Movie duration and gross was converted from object to float. The data frame has no missing or duplicated values.

```
In [28]: movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Movie_Title         250 non-null    object
 1   Year_of_Release     250 non-null    object
 2   Movie_Rating        250 non-null    object
 3   Movie_Duration      250 non-null    float64
 4   Movie_Genre         250 non-null    object
 5   Movie_Votes         250 non-null    int32
 6   Movie_Description   250 non-null    object
 7   Movie_Director      250 non-null    object
 8   Movie_Stars         250 non-null    object
 9   Gross               250 non-null    float64
 10  Metascore_category  250 non-null    category
dtypes: category(1), float64(2), int32(1), object(7)
memory usage: 19.1+ KB
```

```
In [27]: movies.isnull().sum()
```

```
Out[27]: Movie_Title           0
         Year_of_Release       0
         Movie_Rating          0
         Movie_Duration        0
         Movie_Genre           0
         Movie_Votes           0
         Movie_Description     0
         Movie_Director        0
         Movie_Stars           0
         Gross                 0
         Metascore_category    0
         dtype: int64
```

The Final Top 250 IMDB Data Frame

```
In [29]: movies.head(5)
```

Out[29]:

ie_Title	Year_of_Release	Movie_Rating	Movie_Duration	Movie_Genre	Movie_Votes	Movie_Description	Movie_Director	Movie_Stars	Gross	Metascore_category
The wshank emption	1994	9.3	142.0	Drama	2757830	Over the course of several years, two convicts...	Frank Darabont	[Tim Robbins, Morgan Freeman, Bob Gunton, Will...	28.34	80-90
The odfather	1972	9.2	175.0	Crime	1919018	Don Vito Corleone, head of a mafia family, dec...	Francis Ford Coppola	[Marlon Brando, Al Pacino, James Caan, Diane K...	134.97	90-100
he Dark Knight	2008	9.0	152.0	Action	2730773	When the menace known as the Joker wreaks havo...	Christopher Nolan	[Christian Bale, Heath Ledger, Aaron Eckhart, ...	534.86	80-90
indler's List	1993	9.0	195.0	Biography	1389849	In German-occupied Poland during World War II,...	Steven Spielberg	[Liam Neeson, Ralph Fiennes, Ben Kingsley, Car...	96.90	90-100
2 Angry Men	1957	9.0	96.0	Crime	817053	The jury in a New York City murder trial is fr...	Sidney Lumet	[Henry Fonda, Lee J. Cobb, Martin Balsam, John...	4.36	90-100

Data Columns Description:

The dataset consists of the top 250 IMDB movies(rows) with 11 columns in total.

Movie Title: refers to the name of the movies and it serves as a unique identifier for each movie entry in the dataset.

Year of Release: This is the specific year in which each movie was released to the public.

Movie Rating: This attribute represents the rating of the movie by the public.

Movie Duration: This provides information about the length of each movie.

Movie Genre: This represents the genre to which each movie (e.g., "Action," "Drama,", "Crime" etc.).

Movie Votes: This refers to the number of votes received by the movie, and it may indicate the level of audience engagement or popularity of the movie.

Movie Description: This refers to a summary of the movie's plot or storyline.

Movie Directors: These are the names of the individuals responsible for directing the movie.

Movie Stars: This are the names of the prominent cast members in the movie (lead actors/actresses)

Gross: This represents the gross revenue or earnings generated by the movie.

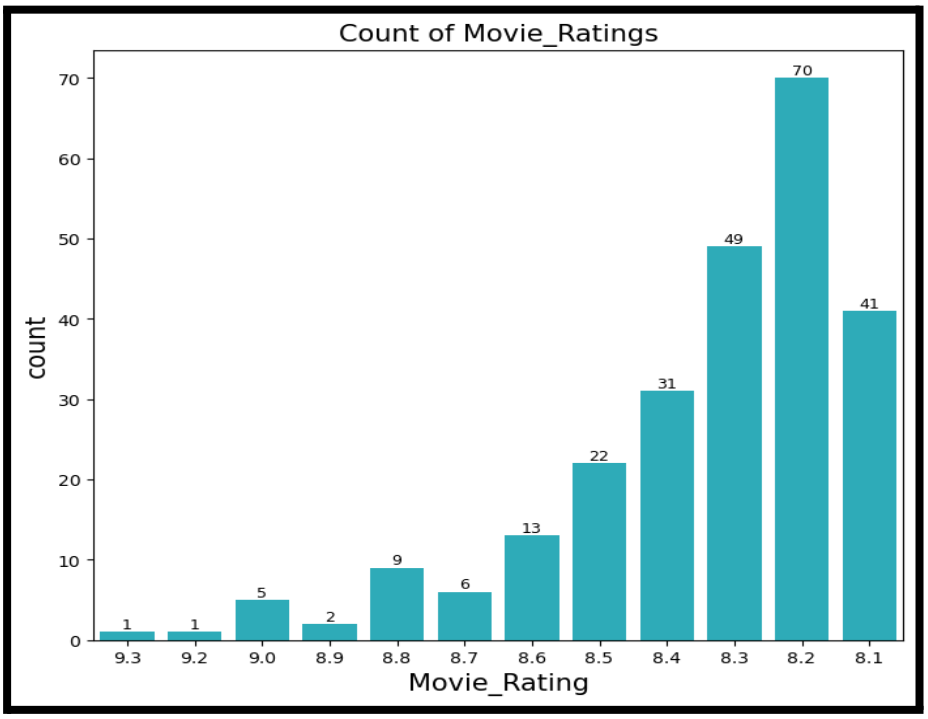
Meta Score category: This represents the categorization of the movie's Meta score.

Exploratory Data Analysis

Univariate Analysis

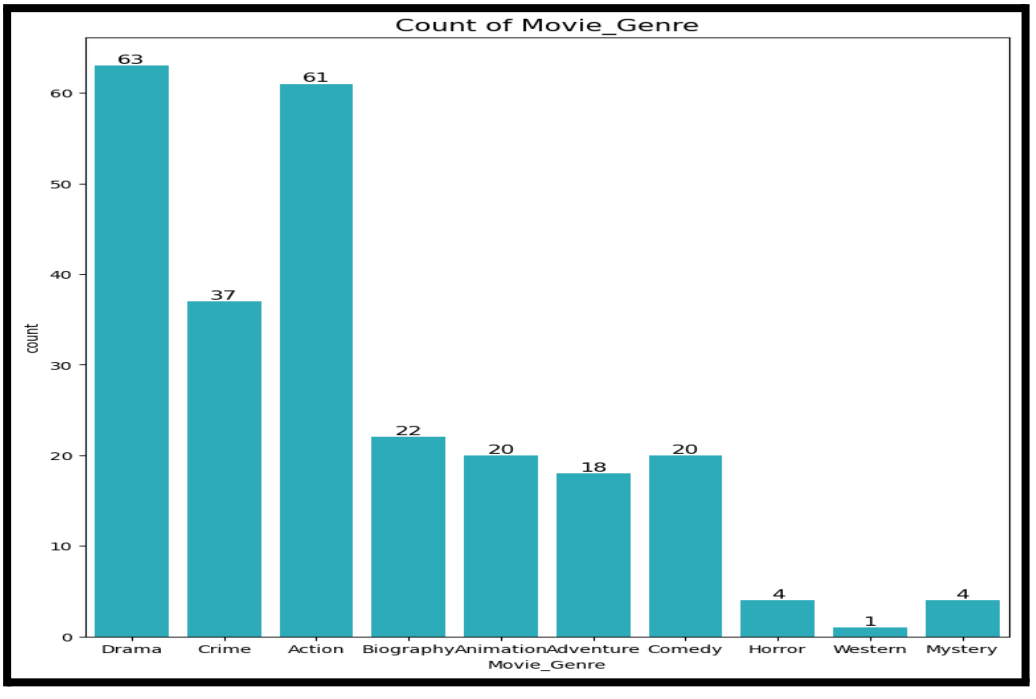
1. Movies Ratings that occurred the most among the Top 250 IMDB.

Observation: Of the top 250 movies scrapped for this project, 8.2rating has the highest rating with 70 movies.



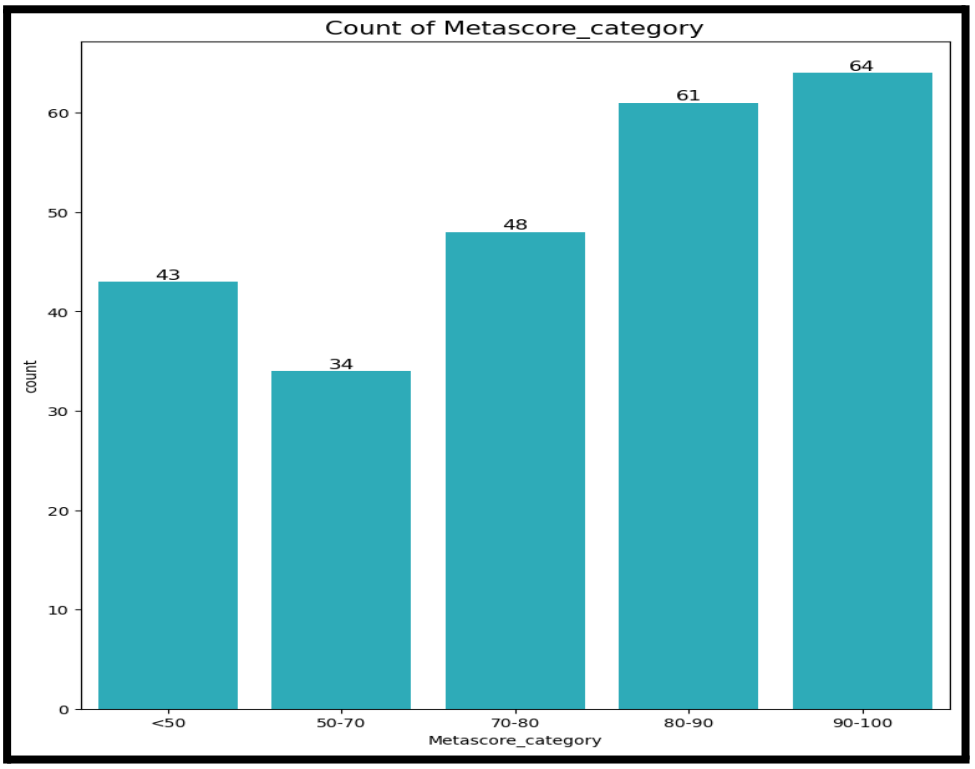
2. Movies Genre that dominated the Top 250 IMDB Movies between 1978 to 2019.

Observation: Drama was the most frequently occurring genre, with 63 films, followed closely by "Action" with 61 movies. This indicates that the Top 250 scraped IMDB dataset contains a significant number of films in the Drama and Action genres. In contrast, there are relatively few movies in Horror, Mystery, and Western genres.



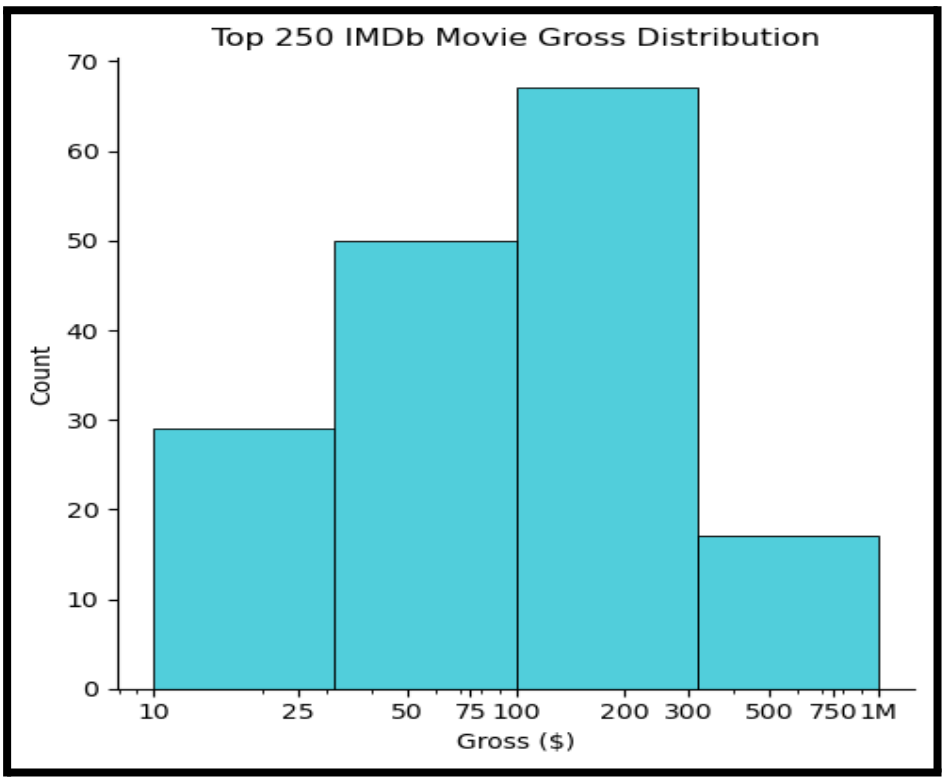
3. MetaScore_Category that has the highest number of Meta scores among the Top 250 movies.

Observation: Among the Top250 movies, 64 movies have a Meta score that falls into the 90-100 score category. This indicates that these movies have received a positive critical reception from the audience. However,43 movies have a <50 Meta score this could be indicative of mixed or negative critical reception.



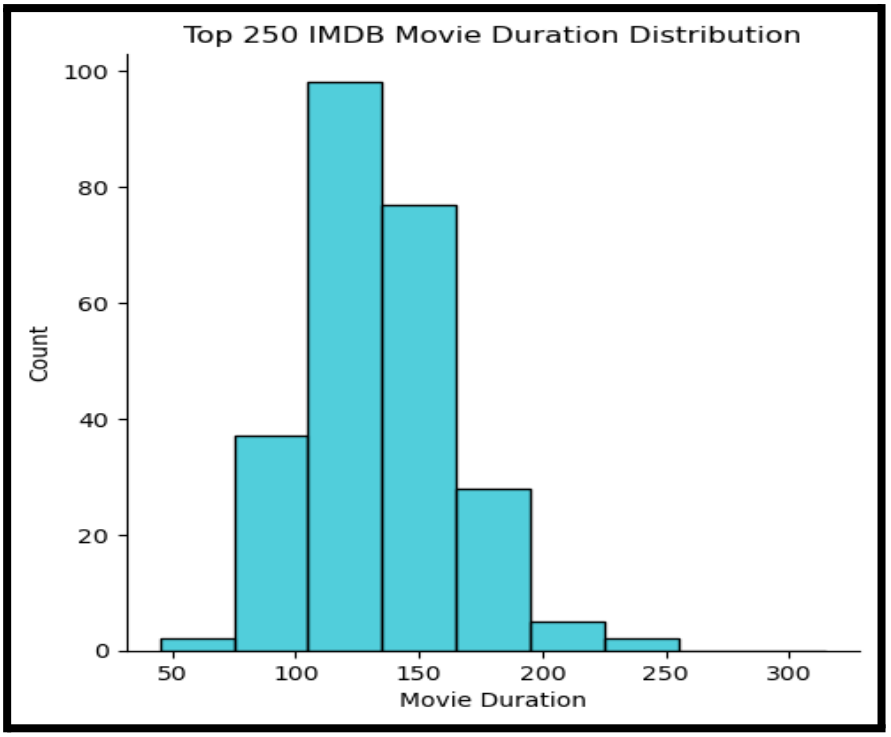
4. Distribution of the Top 250 IMDB Movie Gross

Observation: There was a steady increase arising in the Top 250 IMDB movies gross distribution and we have more of these gross occurring between $100 to $300m and a drop afterwards. However, the maximum gross is $858.37 million. This represents the highest recorded earnings for a movie in the Top 250 dataset, this indicates that the highest-grossing movie made $858.37 million.



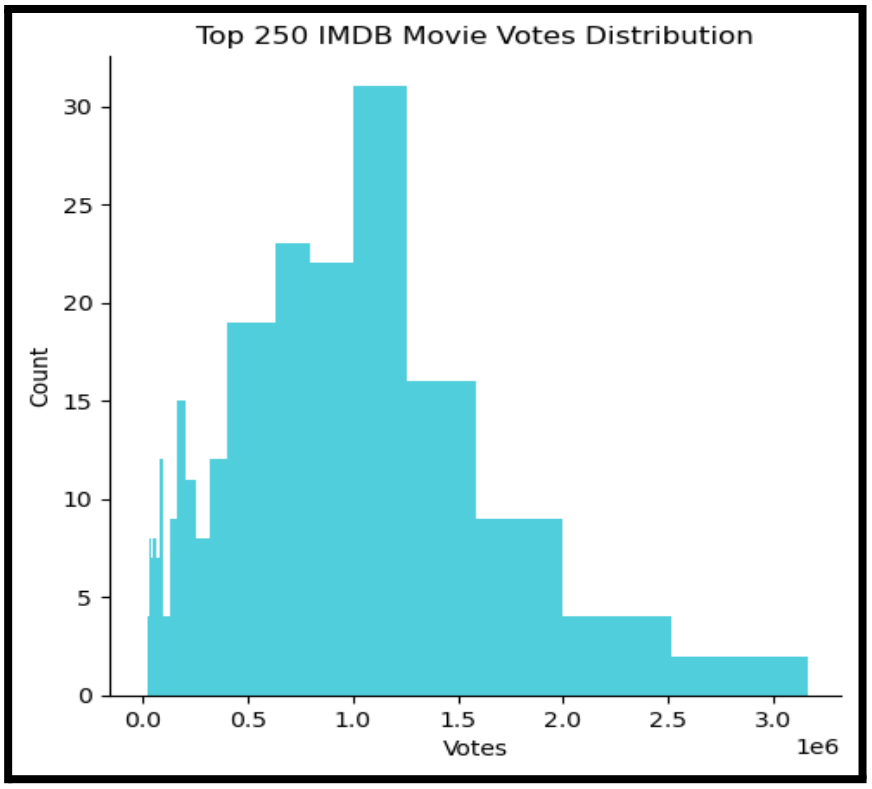
5. Distribution of the Top 250 IMDB Movie Duration

Observation: The Top 250 IMDB movies scraped from the website have a duration of 120 mins, this appears to be sharply 2 hour-long. However, the maximum movie duration is 321 minutes, this means that the longest movie in the dataset has a duration of 321 minutes or approximately 5 hours and 21 minutes.



6. Distribution of the Top 250 IMDB Movie Votes

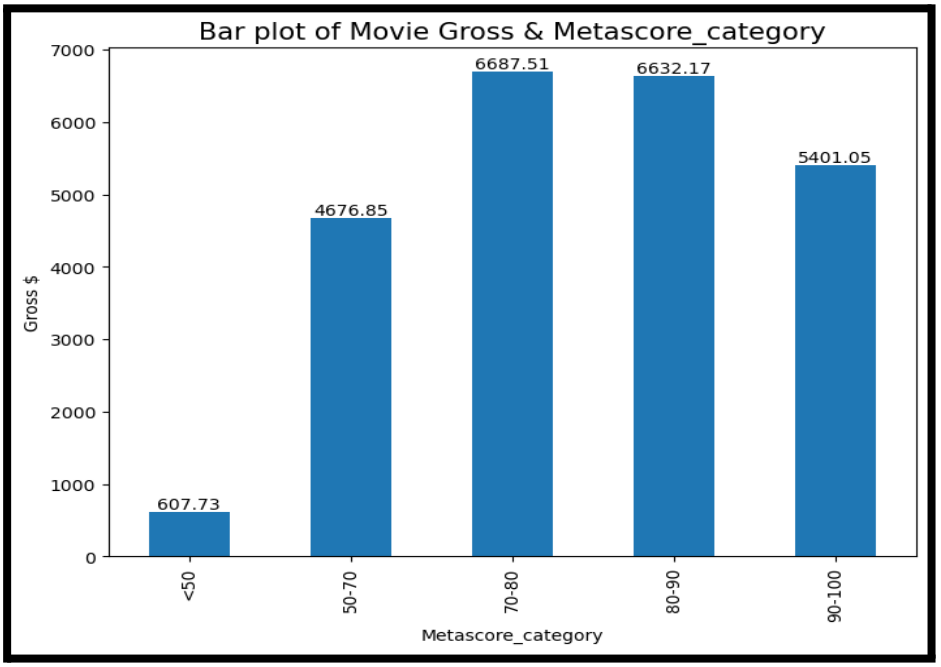
Observation: The maximum number of votes for a movie is 2,756,853. This indicates that the movie with the most votes in the dataset received 2,756,853 votes from the audience.



BIVARIATE ANALYSIS

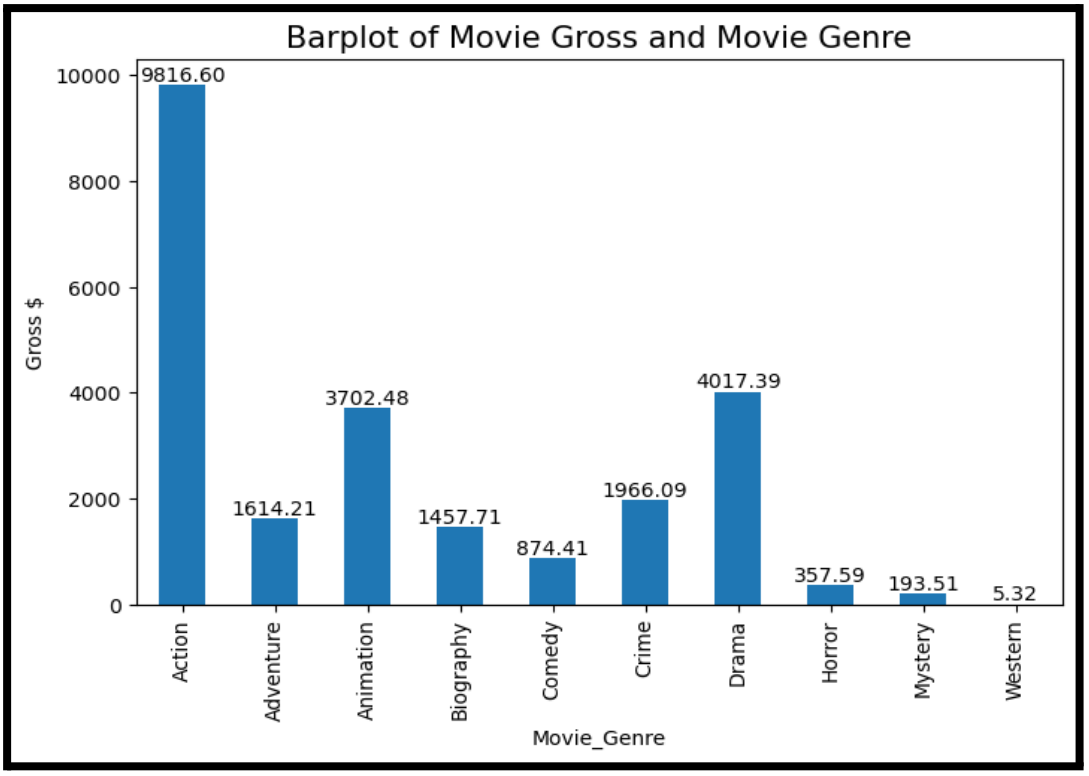
1. Meta score category with the highest Gross

Observation: The bar plot shows that the meta score category between **70-80 scores** has the highest gross of **($6687.51M)** and movies with **< 50 meta score-category** has the lowest gross of **($607.73M).**



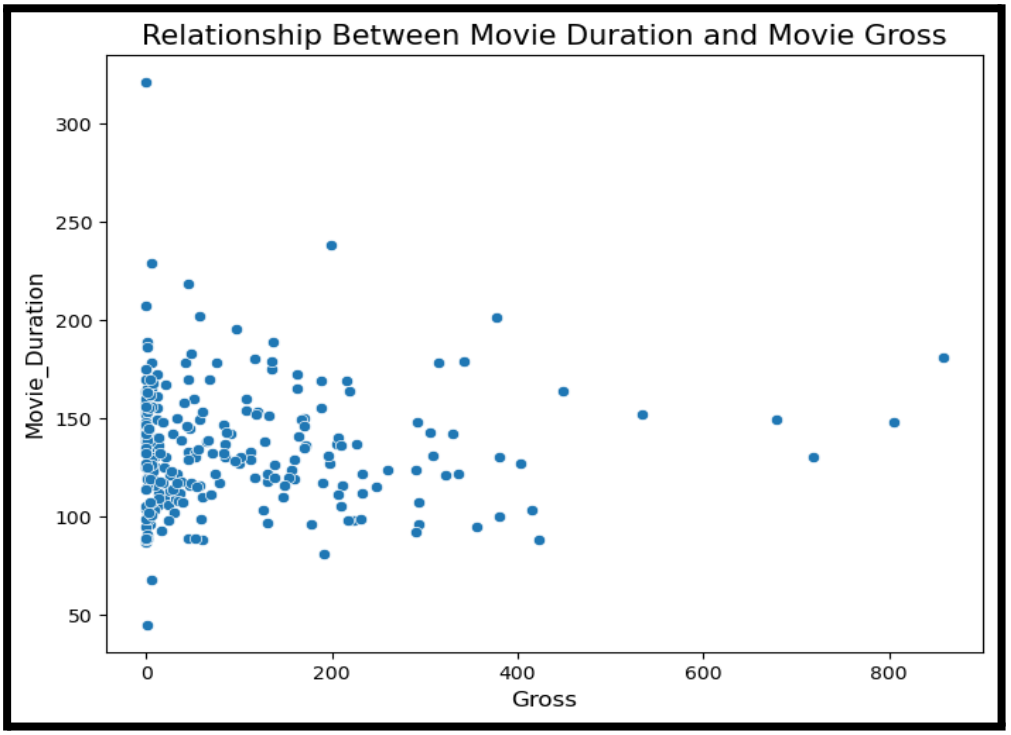
2. Genre of movies that generated the highest gross(revenue) among the Top250 IMDB movies.

Observation: Among the Top250 IMDB movies, the movie genre with the **highest gross** is **Action movies** which has **($9816.60M**) and **Western movies genre** has the **lowest movie gross.**



3. Relationship between Movie Duration and Movies Gross.

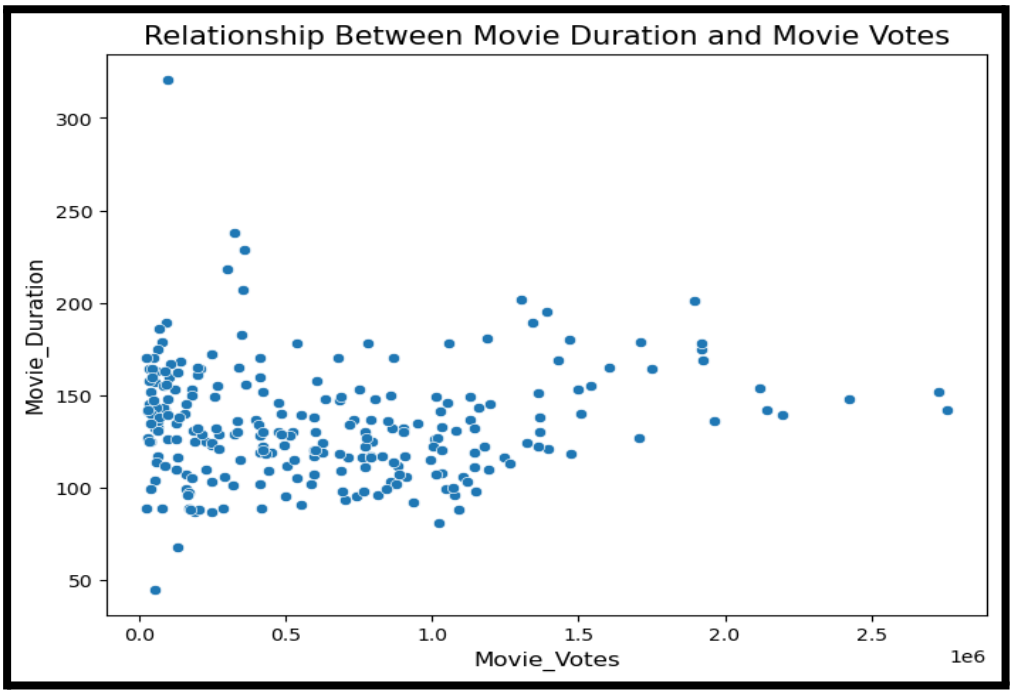
Observation: The scatterplot shows that Movie duration and Gross has a very weak positive correlation with 0.0339 correlation coefficient, and this means that as one variable increases, the other variable tends to increase slightly, but the relationship is not strong.




```
Correlation coefficient:0.033981140174730244
```

4. Relationship between Movies Duration and Movie Votes.

Observation: The scatterplot shows that Movie duration and Movie Votes has a weak positive correlation between the two variables with 0.0524 correlation coefficient and this means that as one variable increases, the other variable tends to increase slightly, but the relationship is not strong.

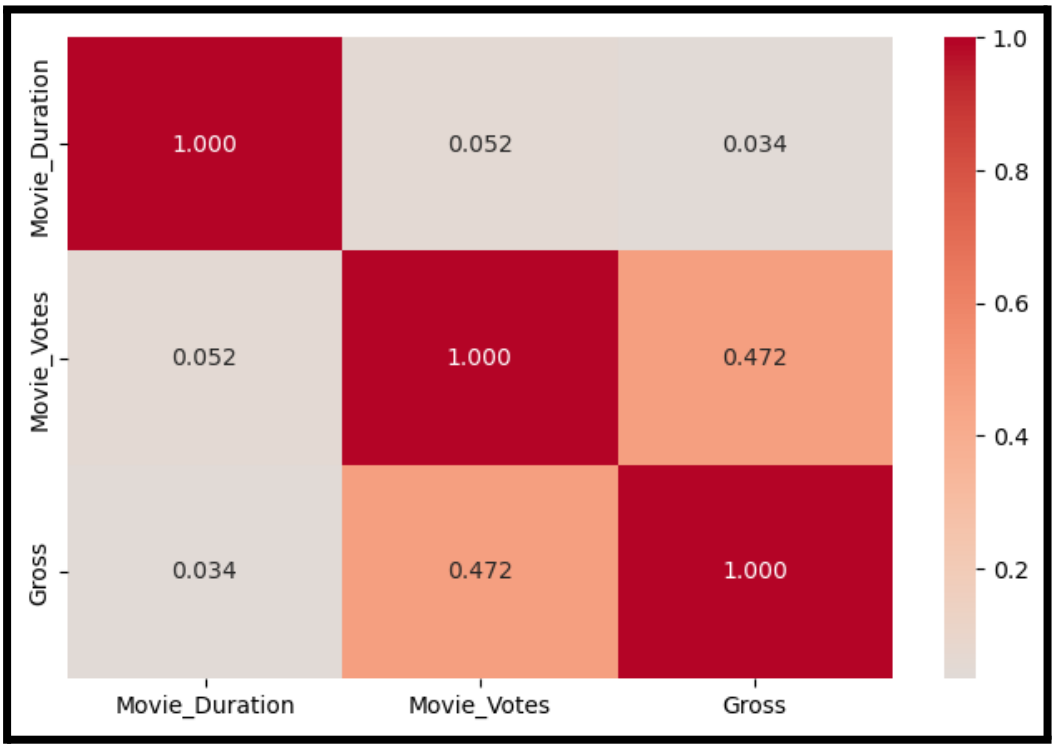



```
Correlation coefficient: 0.05240364392191251
```

MULTIVARIATE ANALYSIS

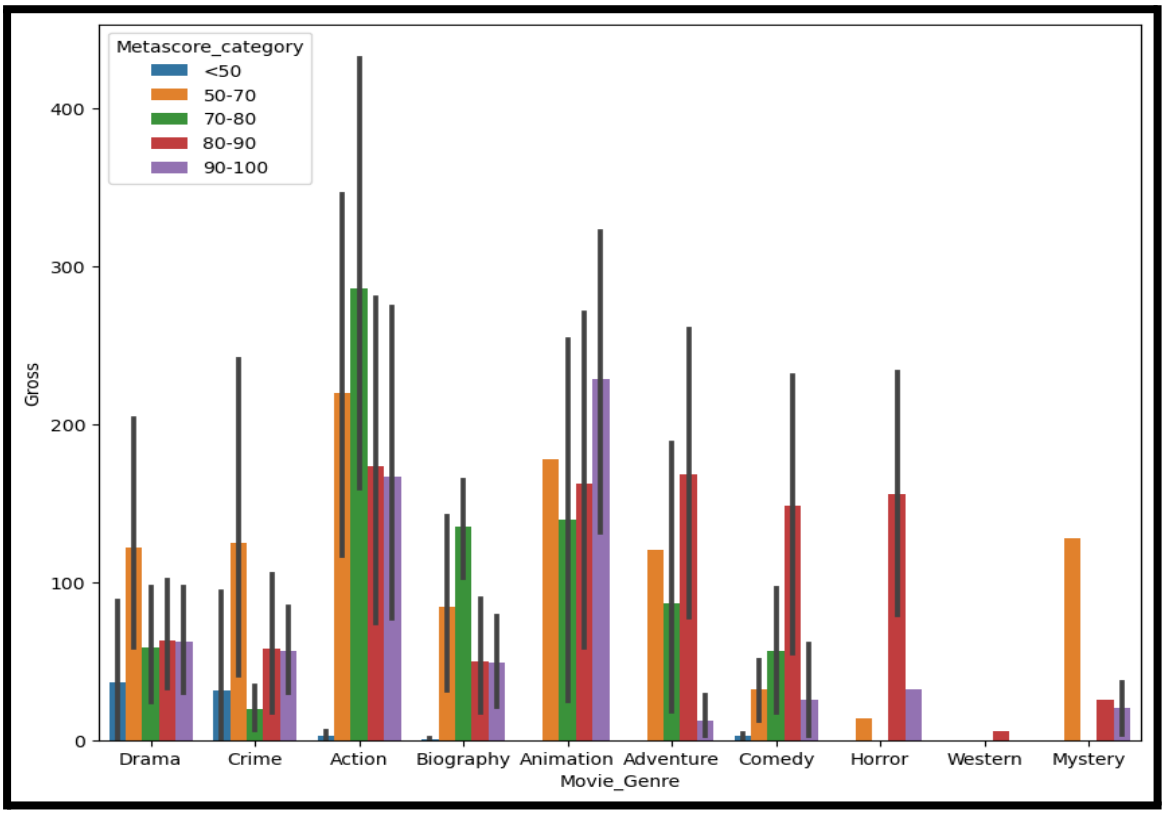
1. Correlation between top 250 movies variables

Observation: From the Top250 IMDB movies, the heatmap shows that Movie Gross has a positive correlation with the Movie Votes and Movie Duration. This means that as Movie Gross increases other variables increase respectively.



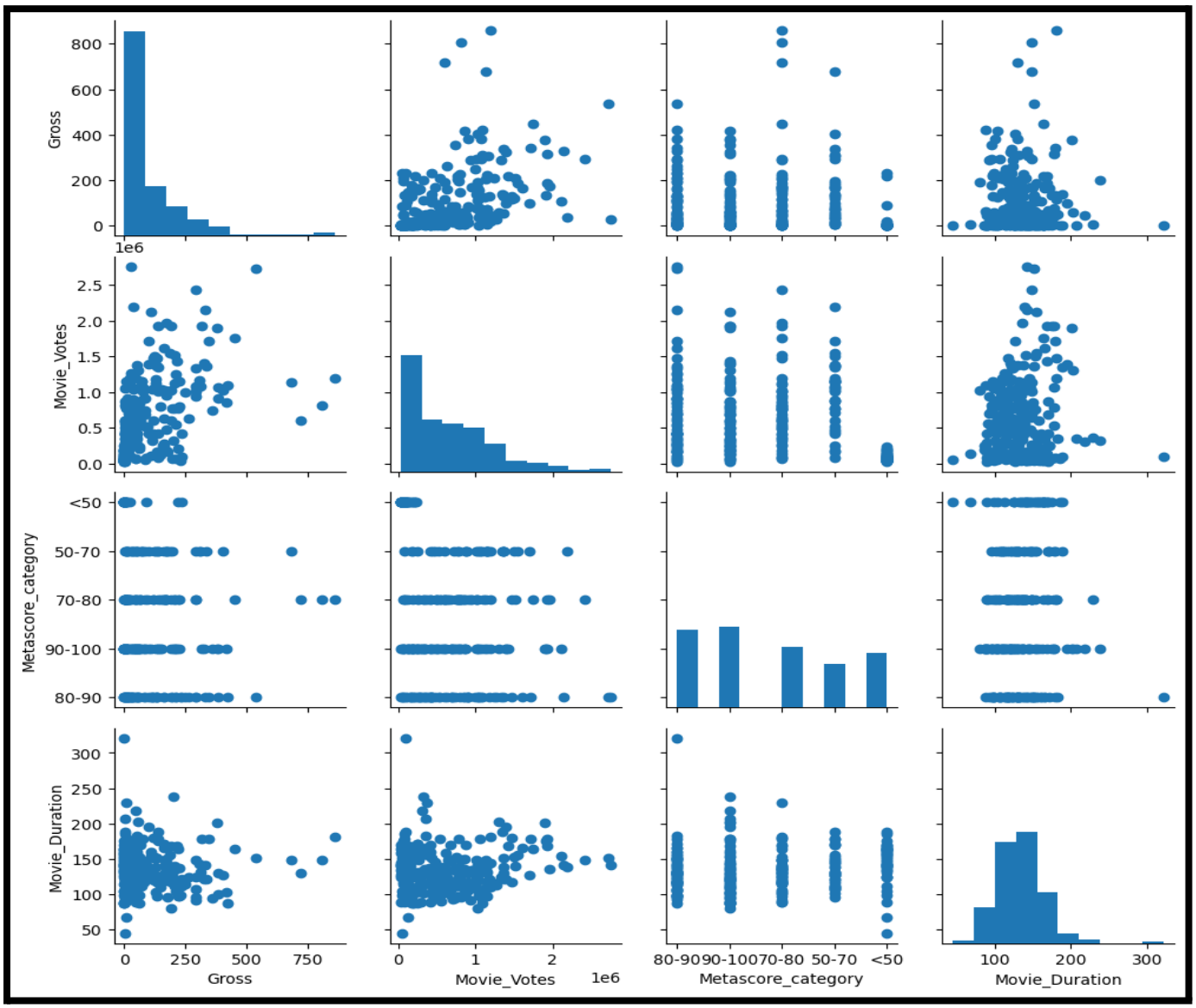
2. The movie Genre with the highest Gross and Meta score category.

Observation: Action movies have higher gross compared to another movie genre and the movie has different meta score categories with 70-80 scores being the highest.



3. Relationship between the four variables (Gross, Movie Votes, Metascore_category and Movie Duration?

Observation: The 4 variables (Gross, Movie Votes, Metascore_category and Movie Duration] are not well associated with one another but they are related to one another in a way. The general pattern between Gross and the 3 distinct dimension is not strong but non-linear because Gross is a measure of revenue. The relationship between Movie duration and Movie Vote are point clustered and there is no significant correlation between them.



CONCLUSION:

The insights from this project will help in understanding audience preferences, analyzing the popularity of different genres and identifying trends in the movie industry and to evaluate movies for critical reception, quality, and rating distribution for comparisons.