

MANAGERIAL REPORT ON

IMDb Sci- Fi Movie Data Set

DEVP-1



SUBMITTED TO: PROF. Amarnath Mitra

BATCH: **BDA-04**
SECTION: **H**

SUBMITTED BY: **ANSH LOOMBA**

ROLL NO. – 045009

Project Objectives:

- General Description of Data
- Analysis : Basic Descriptive & Mathematical or Statistical Analysis
- Findings & Inferences
- Managerial Insights
- Implications

Description of Data :

The data collected by me was for the Sci-Fi movie segment on IMDb (<https://www.imdb.com/search/title/?genres=sci-fi&>), by designing a web scrapper I collected the data in the form a list and then converting the same into a dataframe.

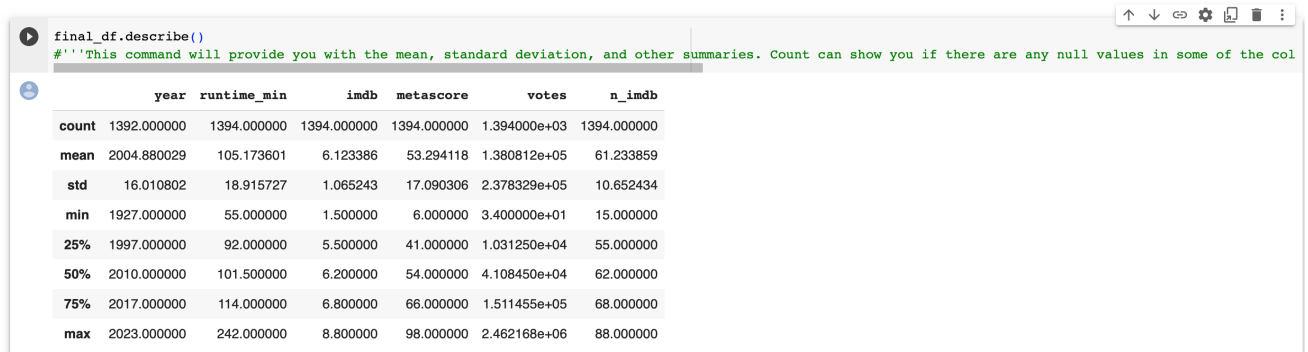
```
[9] final_df.head()
```

	movie	year	rating	genre	runtime_min	imdb	metascore	votes	n_imdb
0	The Flash	2023.0	PG-13	[Action, Adventure, Fantasy]	144	6.8	55	163901	68.0
1	Blue Beetle	2023.0	PG-13	[Action, Adventure, Sci-Fi]	127	6.7	61	27034	67.0
2	Meg 2: The Trench	2023.0	PG-13	[Action, Adventure, Horror]	116	5.2	40	38486	52.0
3	Teenage Mutant Ninja Turtles: Mutant Mayhem	2023.0	PG	[Animation, Action, Adventure]	99	7.4	74	29438	74.0
4	Guardians of the Galaxy Vol. 3	2023.0	PG-13	[Action, Adventure, Comedy]	150	8.0	64	310205	80.0

This command shows you the first 5 rows of your dataframe. It helps you see that nothing looks weird and everything is ready for analysis.

Analysis, Findings & Inferences:

▼ Exploratory Data Analysis

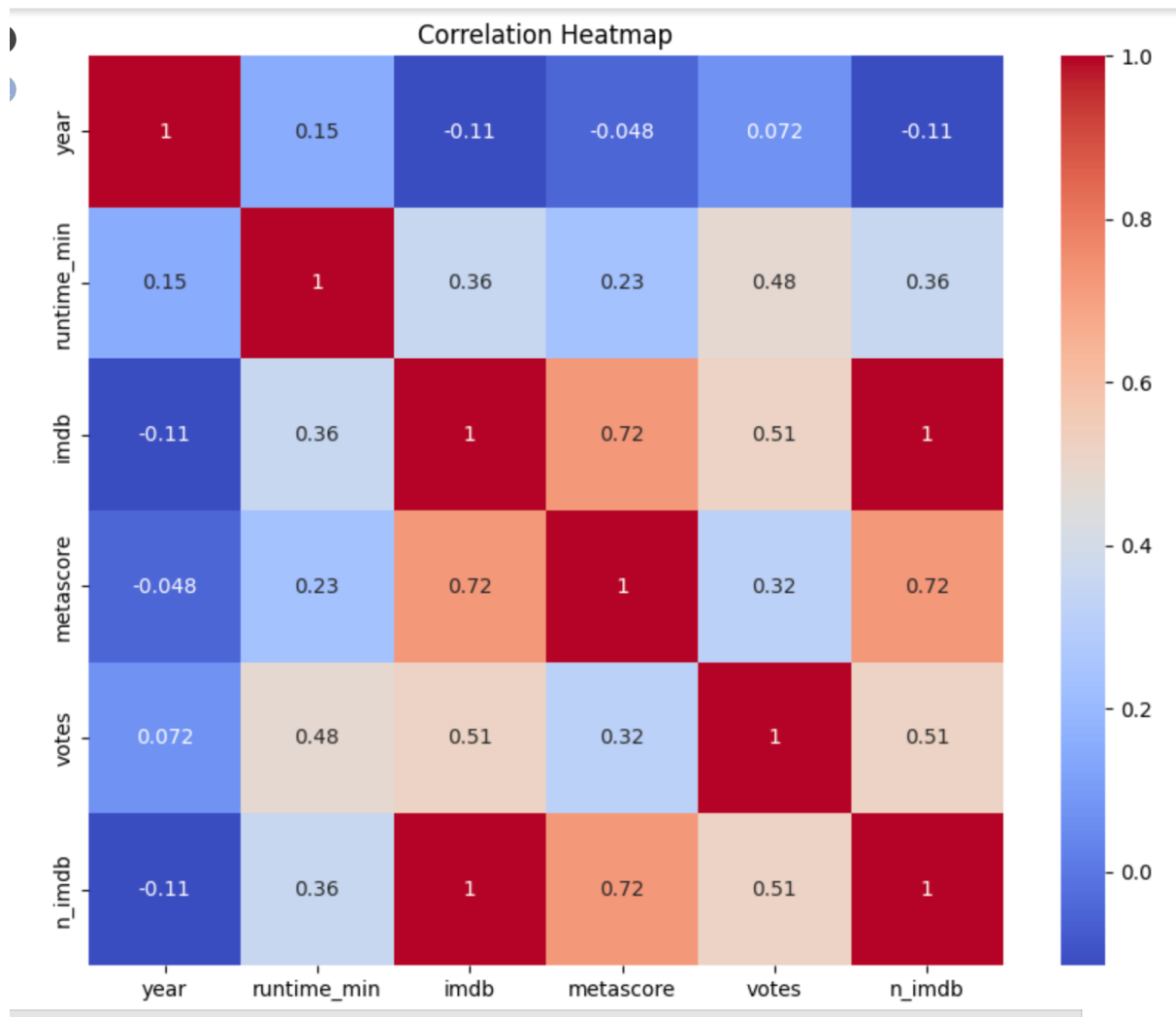


```
final_df.describe()
#'''This command will provide you with the mean, standard deviation, and other summaries. Count can show you if there are any null values in some of the col
```

	year	runtime_min	imdb	metacore	votes	n_imdb
count	1392.000000	1394.000000	1394.000000	1394.000000	1.394000e+03	1394.000000
mean	2004.880029	105.173601	6.123386	53.294118	1.380812e+05	61.233859
std	16.010802	18.915727	1.065243	17.090306	2.378329e+05	10.652434
min	1927.000000	55.000000	1.500000	6.000000	3.400000e+01	15.000000
25%	1997.000000	92.000000	5.500000	41.000000	1.031250e+04	55.000000
50%	2010.000000	101.500000	6.200000	54.000000	4.108450e+04	62.000000
75%	2017.000000	114.000000	6.800000	66.000000	1.511455e+05	68.000000
max	2023.000000	242.000000	8.800000	98.000000	2.462168e+06	88.000000

- 1. Time Period of Movies:** The dataset covers movies released from 1927 to 2023, with an average release year of approximately 2004. This indicates that the dataset is fairly recent and includes movies from a broad historical range.
- 2. Runtime:** The average movie runtime is about 105 minutes, with a minimum of 55 minutes and a maximum of 242 minutes. Most movies (75%) have runtimes between 92 and 114 minutes.
- 3. IMDb Ratings:** The average IMDb rating is around 6.12, with a minimum rating of 1.5 and a maximum rating of 8.8. The ratings exhibit a moderate spread, indicating that movies vary in quality, but the average rating is slightly above average.
- 4. Metascore:** The average Metascore is approximately 53.29, with scores ranging from 6 to 98. The Metascores tend to be lower on average compared to IMDb ratings, indicating potential differences in critical and audience reception.
- 5. Votes:** The number of votes varies widely, with an average of approximately 138,081 votes. The dataset includes movies with as few as 34 votes and as many as 2,462,168 votes, suggesting a diverse set of movies in terms of popularity and audience engagement.

- 6. IMDb vs. Metascore Correlation:** Further analysis is needed to determine the exact correlation coefficient between IMDb ratings and Metascores. A positive correlation would suggest that highly-rated movies on IMDb tend to receive higher Metascores from critics, while a negative correlation would indicate a disconnect between critical and audience opinions.
- 7. IMDb vs. Runtime Correlation:** A correlation analysis can determine if there is a relationship between movie runtime and IMDb ratings. A positive correlation would suggest that longer movies tend to receive higher IMDb ratings, while a negative correlation would indicate the opposite.
- 8. Metascore vs. Runtime Correlation:** A correlation analysis can also determine if there is a relationship between movie runtime and Metascores. A positive correlation would imply that longer movies tend to receive higher Metascores from critics, while a negative correlation would suggest the opposite.
- 9. Popularity vs. Ratings:** Analyzing the correlation between the number of votes a movie receives and its IMDb rating or Metascore can reveal whether more popular movies tend to receive higher ratings or critical acclaim.

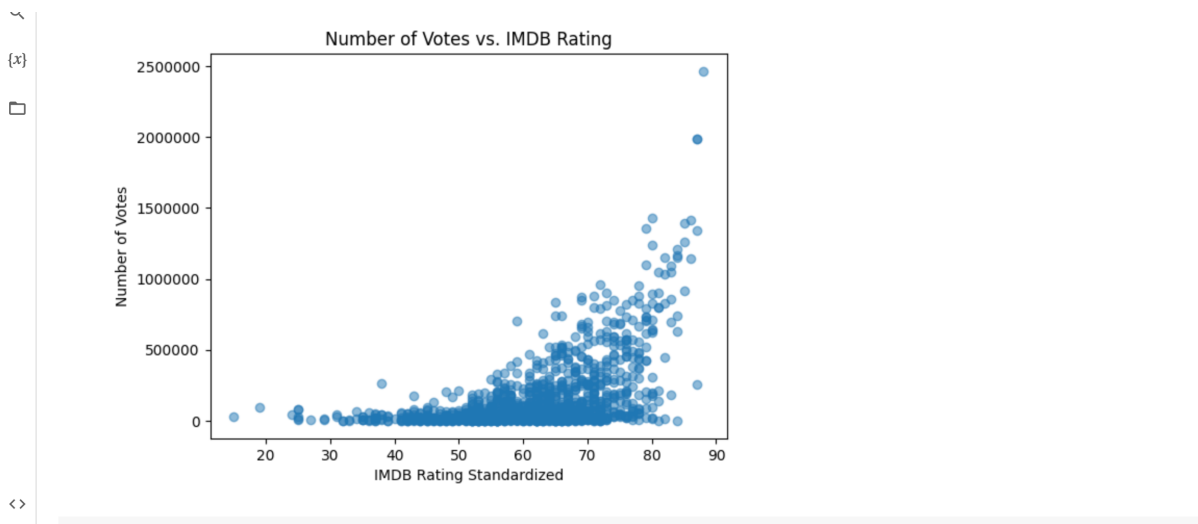


We can see that the strongest correlation is between the IMDB score and the metacore. This is not surprising since it's likely that two movie rating systems rate similarly.

The next strongest correlation we can see is between the IMDB rating and the number of votes. This is interesting because as the number of votes increases, you have a more representative sample of the population rating. It's strange to see that there is a weak association between the two, though.

The number of votes roughly increases as the runtime increases as well.

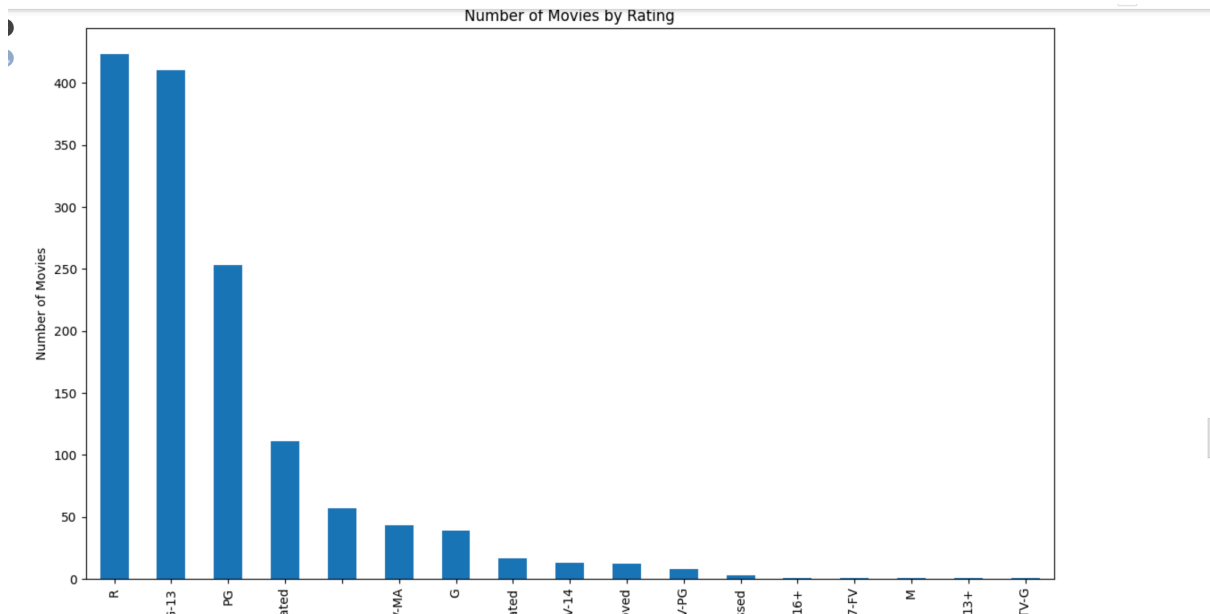
We can also see a slight negative association between IMDB or metascore and the year the movie came out.



IMDB Ratings vs. the Number of Votes

The association above shows some outliers. Generally, we see a greater number of votes on movies that have an IMDB rating of 85 or more. There are fewer reviews on movies with a rating of 75 or less.

Another thing that might be an insight is to see how many movies of each rating there are. This can show you where Sci-Fi tends to land in the ratings data.



We can see that there were a few movies rated as “Approved” and was curious what that was. You can filter down the dataframe with this code to drill down into that:

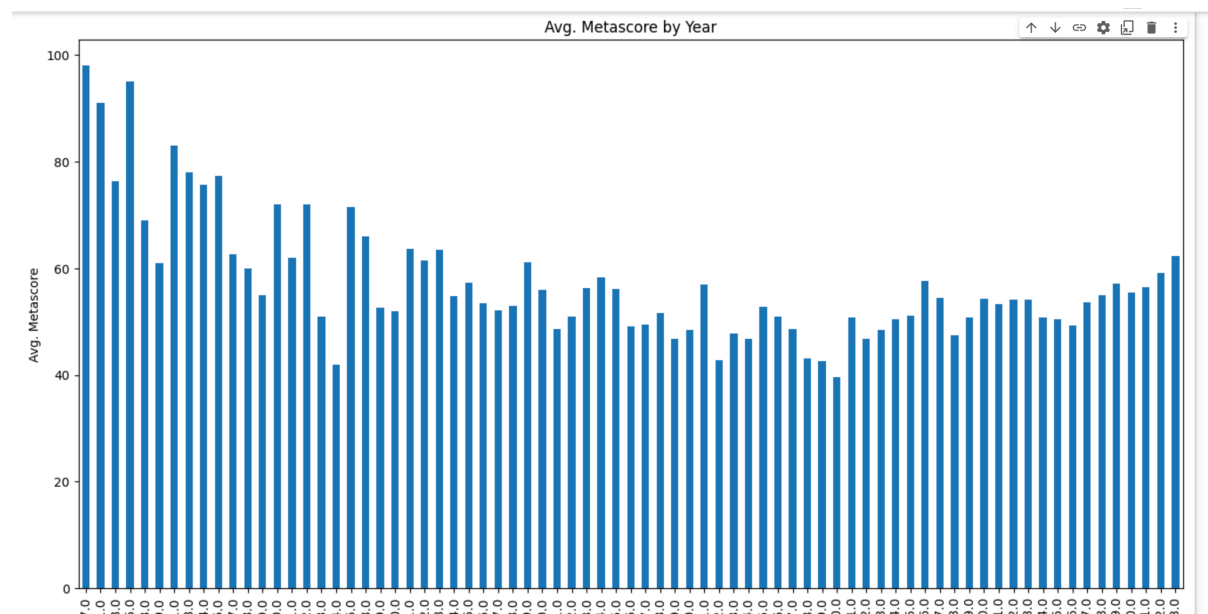
```
[15] final_df[final_df['rating'] == 'Approved']
```

	movie	year	rating	genre	runtime_min	imdb	metascore	votes	n_imdb
315	Barbarella	1968.0	Approved	[Adventure, Comedy, Fantasy]	98	5.8	51	36725	58.0
676	Rollerball	1975.0	Approved	[Action, Sci-Fi, Sport]	125	6.6	56	27590	66.0
787	The Blob	1958.0	Approved	[Horror, Sci-Fi]	86	6.3	58	28268	63.0
807	Invasion of the Body Snatchers	1956.0	Approved	[Drama, Horror, Sci-Fi]	80	7.7	92	52901	77.0
882	The Curse of Frankenstein	1957.0	Approved	[Horror, Sci-Fi, Thriller]	82	7.0	59	12066	70.0
927	On the Beach	1959.0	Approved	[Drama, Romance, Sci-Fi]	134	7.1	55	14022	71.0
948	Demon Seed	1977.0	Approved	[Horror, Sci-Fi]	94	6.3	55	9667	63.0
1067	The Absent Minded Professor	1961.0	Approved	[Comedy, Family, Sci-Fi]	92	6.7	75	8944	67.0
1220	The Damned	1962.0	Approved	[Drama, Fantasy, Horror]	87	6.6	72	3820	66.0
1254	Mighty Joe Young	1949.0	Approved	[Adventure, Drama, Fantasy]	94	7.0	61	5557	70.0
1296	Children of the Damned	1964.0	Approved	[Drama, Horror, Mystery]	89	6.2	42	4852	62.0
1352	The Mouse on the Moon	1963.0	Approved	[Comedy, Romance, Sci-Fi]	82	6.3	62	1867	63.0

This revealed that most of these movies were made before the 80s

We could also check out if any years or decades outperformed others on reviews. I took the average metascore by year and plotted that with the following code to explore further

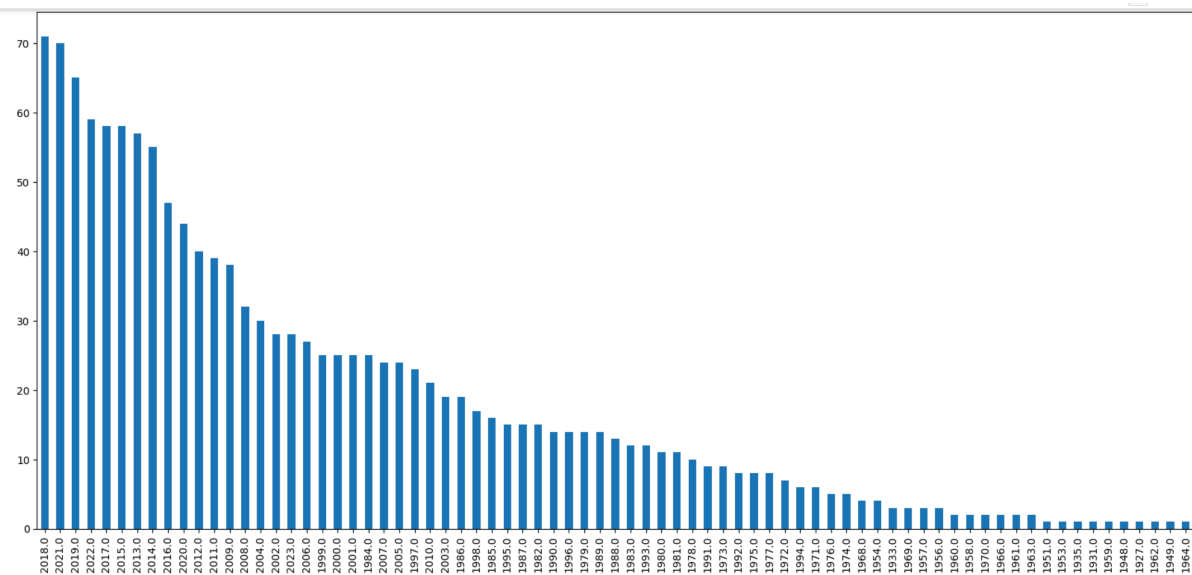
```
# What are the average metascores by year?
final_df.groupby('year')['metascore'].mean().plot(kind='bar',
figsize=(16,8), title="Avg. Metascore by Year", xlabel="Year",
ylabel="Avg. Metascore")
plt.xticks(rotation=90)
plt.plot();
```



There is a gradual, mild decline as you progress through history in the average metascore variable. It seems that ratings have leveled out around 55-60 in the last couple decades. This might be because we have more data on newer movies or newer movies tend to get reviewed more.


```
final_df['year'].value_counts().plot(kind='bar', figsize=[20,9])
```

After running the above code we can see that the 1927 movie only had a sample of 1 review. That score is then biased and over-inflated. We can see that the more recent movies are better represented in reviews.



Managerial or Key Insights of analysing the data:

Analyzing movie data like the one you provided can offer valuable insights for various stakeholders in the film industry, including producers, distributors, and marketers. Here are some key managerial insights and analyses that can be derived from this dataset:

- 1. Genre Popularity:** You can analyze which genres are currently popular among moviegoers. For example, you can calculate the average ratings and box office performance of movies in different genres to identify trends.
 - 2. Box Office Performance:** Investigate the relationship between a movie's runtime and its box office success. Do longer or shorter movies tend to perform better at the box office?
 - 3. Audience Reception:** Examine how IMDb ratings correlate with Metascore ratings. Are movies that are well-received by critics also favored by audiences? This can help producers and distributors understand the importance of critical reviews in driving audience interest.
 - 4. Franchise Success:** If there are any movie franchises in the dataset (e.g., "Guardians of the Galaxy Vol. 3"), analyze how each installment performed compared to its predecessors. This can provide insights into the franchise's longevity and audience loyalty.
 - 5. Demographic Analysis:** Consider the PG and PG-13 ratings. Analyze whether movies with different ratings have distinct audience demographics. This information can help marketers target specific age groups more effectively.
 - 6. Release Year Trends:** Look at how movie ratings and box office earnings have changed over the years. Are there any trends indicating shifts in audience preferences or industry dynamics?
 - 7. Outliers Identification:** Identify outliers in the dataset. For instance, "Teenage Mutant Ninja Turtles: Mutant Mayhem" has a notably high Metascore compared to the others. Investigate the reasons behind such outliers, as they may provide insights into unexpected successes or failures.
- Marketing Strategies: Study

the relationship between the number of votes and box office earnings. This can help in assessing the effectiveness of marketing and promotional campaigns in generating audience interest and driving box office revenue.

8. **Niche Genres:** Analyze less common genres (e.g., "Animation" in the case of "Teenage Mutant Ninja Turtles: Mutant Mayhem"). Are there niche genres that consistently perform well, indicating potential untapped markets?
9. **Budget Analysis:** If available, incorporate budget data to assess the return on investment (ROI) for each movie. This can provide valuable insights into cost-effective filmmaking strategies.

Implications :

The implications of the insights drawn from the movie dataset can have significant consequences for various stakeholders in the film industry. Here are some implications based on the analyses:

1. **Genre Selection and Investment:** Producers and studios can use genre popularity data to make informed decisions about which types of movies to invest in. If certain genres consistently perform well, they may choose to allocate more resources to produce films in those genres.
2. **Runtime Optimization:** Filmmakers can use insights about the relationship between runtime and box office performance to make decisions about the ideal length of their movies. This can impact the pacing of the story and audience engagement.
3. **Marketing and Promotion Strategies:** Movie marketers can focus their efforts on promoting films that have a high likelihood of success, based on their genre, rating, and expected audience demographics. They can also prioritize marketing campaigns for movies with strong critical acclaim.
4. **Franchise Planning:** Studios can make strategic decisions about the continuation of movie franchises based on the performance of previous installments. If a franchise consistently performs well, they may plan for more sequels or spin-offs.

- 5. Critical Reviews vs. Audience Reception:** Filmmakers and studios can assess the importance of critical reviews versus audience ratings. If there is a strong correlation between the two, they may prioritize efforts to receive positive critical reviews.
- 6. Demographic Targeting:** Marketing teams can tailor their campaigns to specific age groups based on movie ratings. For example, PG-rated movies may be marketed more towards families and children, while PG-13-rated movies may target teenagers and young adults.
- 7. Long-Term Industry Trends:** Industry professionals can monitor the trends in movie ratings, box office earnings, and audience reception over the years. This information can be used for long-term strategic planning and adapting to changes in audience preferences.
- 8. Outlier Exploration:** Understanding outliers can help studios investigate the reasons behind exceptional success or failure. This can lead to valuable lessons and insights for future movie production and marketing strategies.
- 9. Niche Genre Exploration:** Studios may explore niche genres that consistently perform well, potentially tapping into underserved markets and diversifying their film portfolios.
- 10. ROI Assessment:** Incorporating budget data into the analysis can help studios assess the profitability of their movies. This can lead to more informed decisions regarding budget allocation and cost-effective filmmaking practices.

In summary, the implications of analyzing movie data are multifaceted and can impact decisions related to genre selection, marketing strategies, franchise planning, and more. These insights can help the film industry make data-driven decisions to improve the success and profitability of their productions.