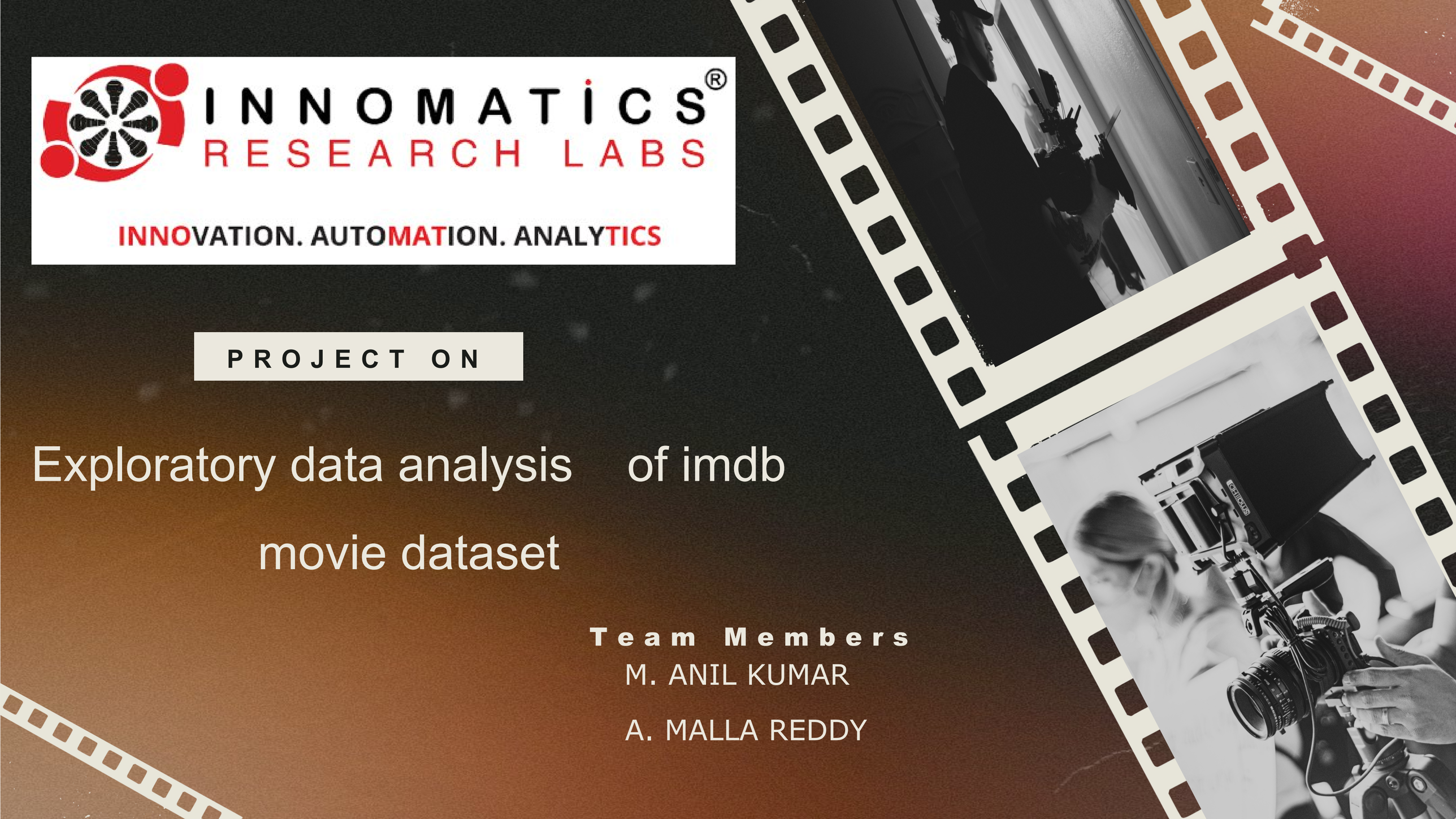
INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory data analysis of imdb movie dataset



Team Members

M. ANIL KUMAR

A. MALLA REDDY

Problem statement

- It is difficult to understand which movies perform well and why.
- Movies are released in different genres, languages, and regions, and each movie has different ratings, runtime, actors and revenue.
- So, we need to analyze IMDB movie data to identify trends and understand what factors make a movie successful.

Objectives

- To find which genres are most popular.
- To analyze movie ratings and revenue.
- To compare regions like Hollywood, Bollywood and Tollywood
- To observe how many movies release every year.
- To study movie runtime and audience preference.
- To understand relationships between votes, revenue and ratings.

Tools Used in This Project

Python (3.12) + Jupyter Notebook

Python Libraries:

- **For Data Extraction:**
  - Beautifulsoup (Web Scraping)
  - re (Regular Expressions)
  - Pandas
  - NumPy
- **For Data Analysis :**
  - Pandas
  - NumPy
- **For Data Visualization :**
  - Matplotlib
  - Seaborn

DATASET OVERVIEW

DATASET SOURCE

Scraped data from IMDB.com using BeautifulSoup (web scraping)

DATASET SHAPE

500 Rows x 12 Columns

KEY COLUMNS

Title, Year, Genre, Rating, Runtime, Votes, Gross, Region, Certificate, Language, Actor, Director

DATA ANALYZING

COLUMN CLEANING

- Column name : Runtime

  The **Runtime** column originally contained values like “2h 8min”.These were converted into total minutes
  (e.g., “2h 8min” -> 128 minutes) for numerical analysis.

- Column name : Rating

  The **Rating** column contained values such as “59% (69)7.1 (2k)100%”.We used regular expressions to extract only the numeric part (e.g., 59% (69)7.1 (2k)100% -> 7.1), and convert the column to numerical analysis

- Column name : Title

  The Title column contained titles with release years eg., “Baahubali (2024)”.The year was extracted and stored
  In a separate column Year (e.g., “Baahubali (2024)” -> 2024), and convert the column to numerical analysis

CHECK NULL VALUES

```
df.isna().sum()
```

```
Title           0
Year            0
Genre           0
Rating          0
Runtime         0
Votes           0
Gross_USA       0
Region          0
Certificate     0
Language        0
Actor           0
Director        0
dtype: int64
```

Summery of DataFrame

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Title        500 non-null    object 
 1   Year         500 non-null    int64  
 2   Genre        500 non-null    object 
 3   Rating       500 non-null    float64
 4   Runtime      500 non-null    int64  
 5   Votes        500 non-null    int64  
 6   Gross_USA    500 non-null    float64
 7   Region       500 non-null    object 
 8   Certificate  500 non-null    object 
 9   Language     500 non-null    object 
 10  Actor        500 non-null    object 
 11  Director     500 non-null    object 
dtypes: float64(2), int64(3), object(7)
memory usage: 47.0+ KB
```

Statistics of Numeric Columns

```
df.describe()
```

	Year	Rating	Runtime	Votes	Gross_USA
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	2006.874000	7.380800	127.926000	95230.764000	458.105380
std	9.981653	1.256865	41.121752	57613.939989	254.454268
min	1990.000000	5.000000	60.000000	1188.000000	1.420000
25%	1998.000000	6.400000	91.750000	48763.750000	239.827500
50%	2007.000000	7.500000	128.000000	91690.000000	462.485000
75%	2015.000000	8.400000	164.000000	144010.000000	670.350000
max	2024.000000	9.500000	200.000000	199732.000000	898.960000

First 5 Records

```
df.head()
```

	Title	Year	Genre	Rating	Runtime	Votes	Gross_USA	Region	Certificate	Language	Actor	Director
0	Baahubali 213	2006	Biography	7.7	172	130220	445.91	Hollywood	UA	Hindi	Ram Charan	Boyapati
1	Avatar 738	2000	Biography	7.6	146	57469	645.98	Tollywood	UA	English	Mahesh Babu	Bong Joon-ho
2	Interstellar 207	1999	Comedy	8.8	69	169784	655.96	Tollywood	A	Telugu	Leonardo DiCaprio	Christopher Nolan
3	Vikram 313	2001	Sci-Fi	7.5	121	86151	113.82	Bollywood	UA	Telugu	Vijay	Lokesh Kanagaraj
4	Titanic 922	2022	Adventure	9.1	131	159672	369.34	Hollywood	A	Hindi	Ram Charan	Lokesh Kanagaraj

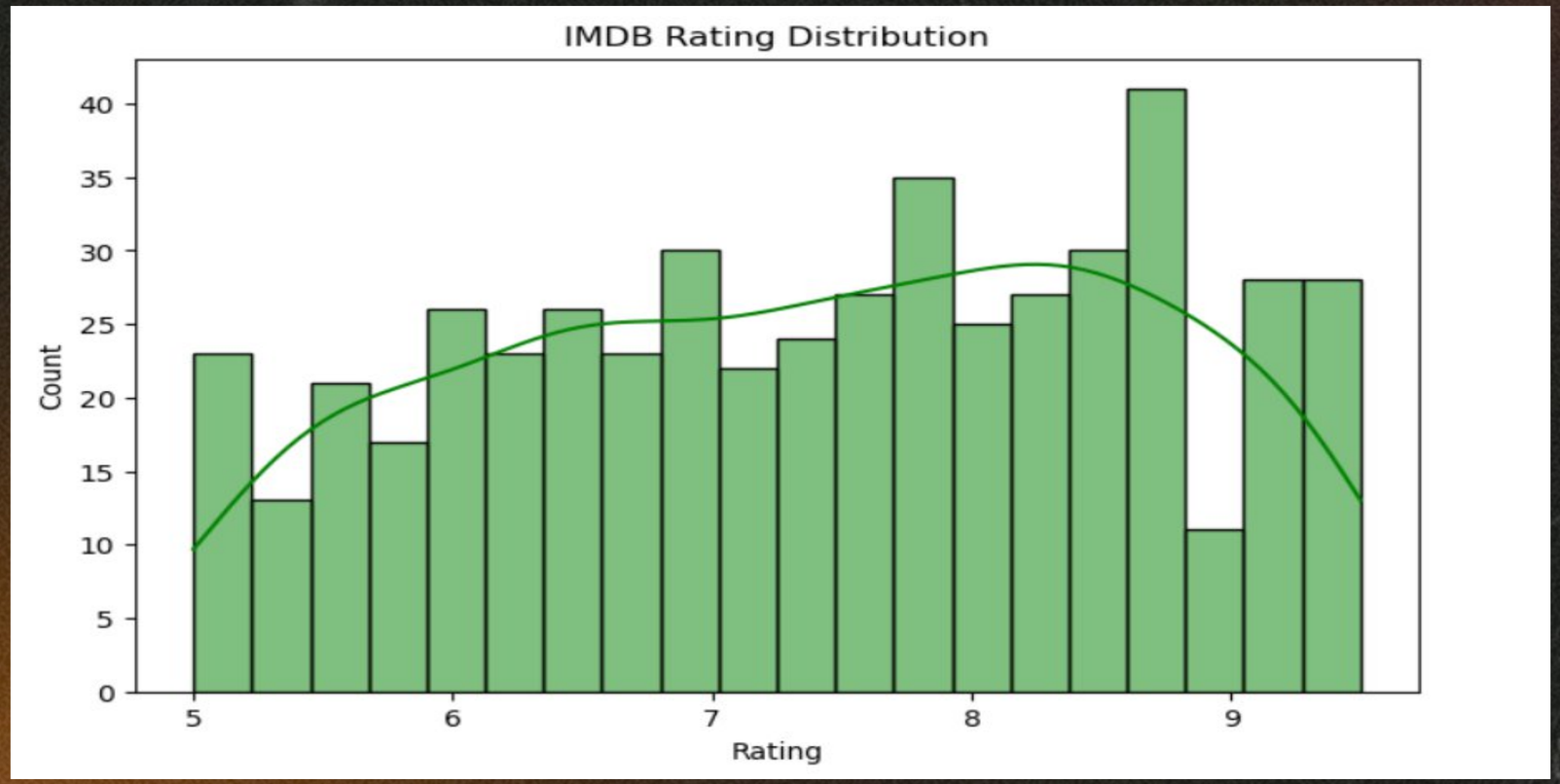
Last 5 Records

```
df.tail()
```

	Title	Year	Genre	Rating	Runtime	Votes	Gross_USA	Region	Certificate	Language	Actor	Director
495	Pathaan 505	1990	Thriller	8.9	75	104240	158.96	Bollywood	U	Telugu	Prabhas	Martin Scorsese
496	Iron Man 122	2002	Thriller	5.4	172	108818	698.86	Hollywood	A	Hindi	Allu Arjun	Lokesh Kanagaraj
497	Baahubali 230	2021	Sci-Fi	8.2	129	145993	752.99	Tollywood	A	English	Yash	Christopher Nolan
498	Titanic 993	2002	Comedy	5.9	196	171162	783.64	Tollywood	UA	Hindi	Shah Rukh Khan	Boyapati
499	Dangal 649	1998	Thriller	6.8	92	100185	721.06	Tollywood	A	English	Mahesh Babu	Sandeep Reddy

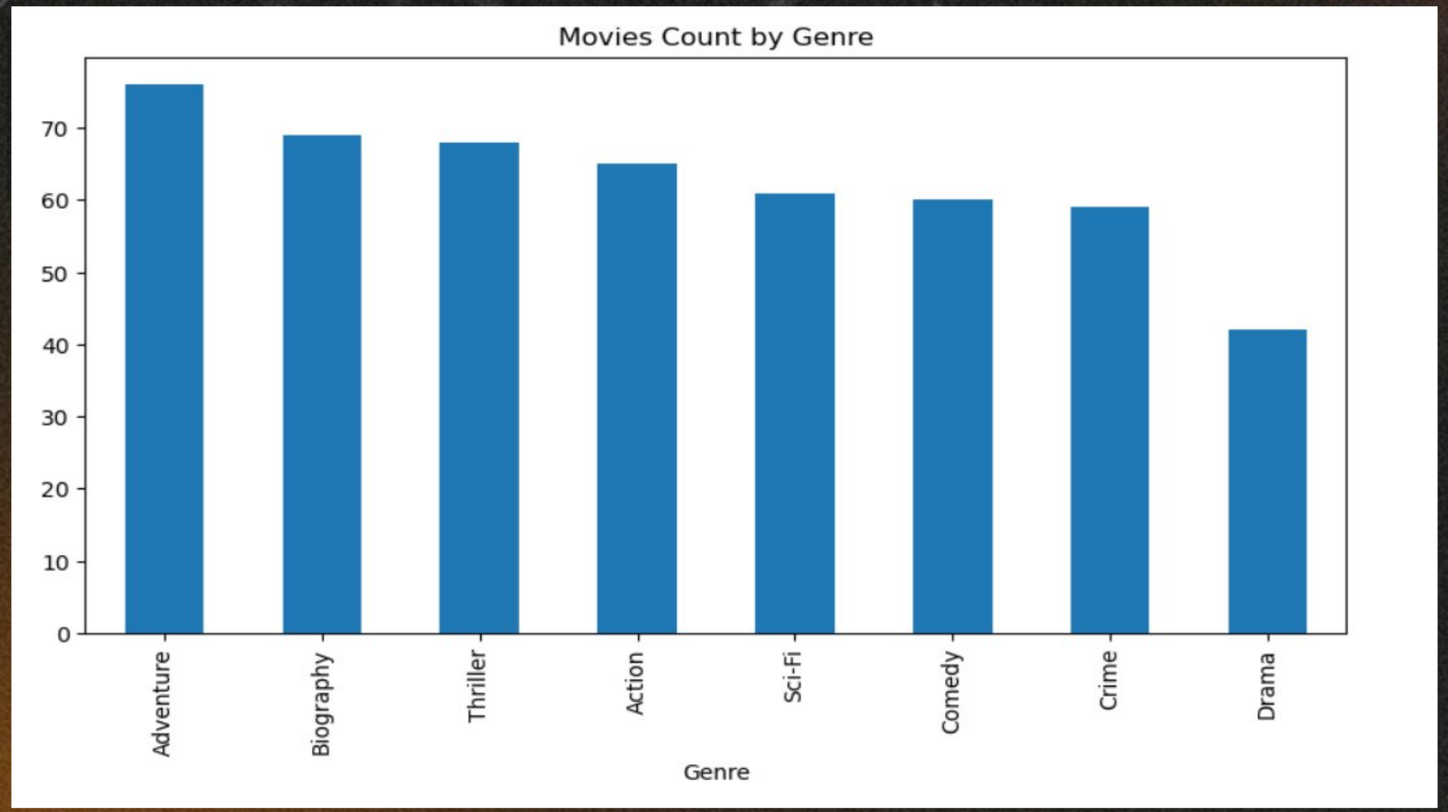
DATA VISUALIZATION

UNIVARIATE ANALYSIS :



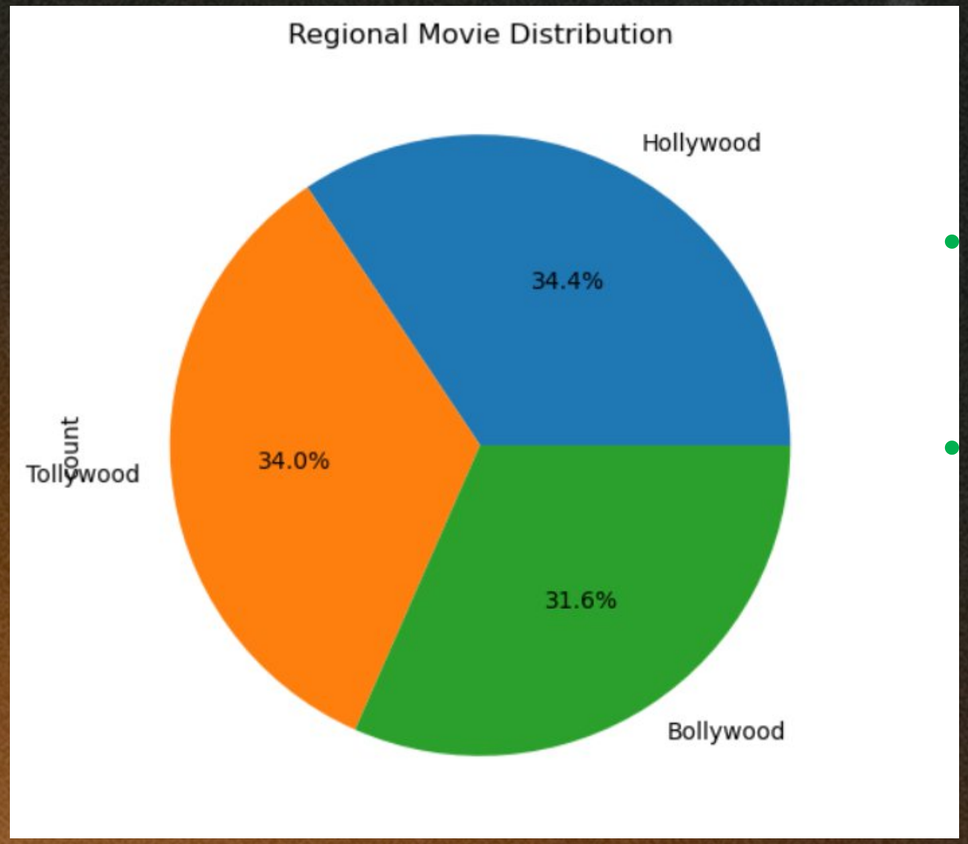
- From the rating distribution graph, we observe that most movies are rated between **6.5 and 8.5**.
- This means the majority of movies are positively rated by the audience.
- Very high and very low ratings are rare, indicating general audience acceptance of most films.

Genre-wise movie count



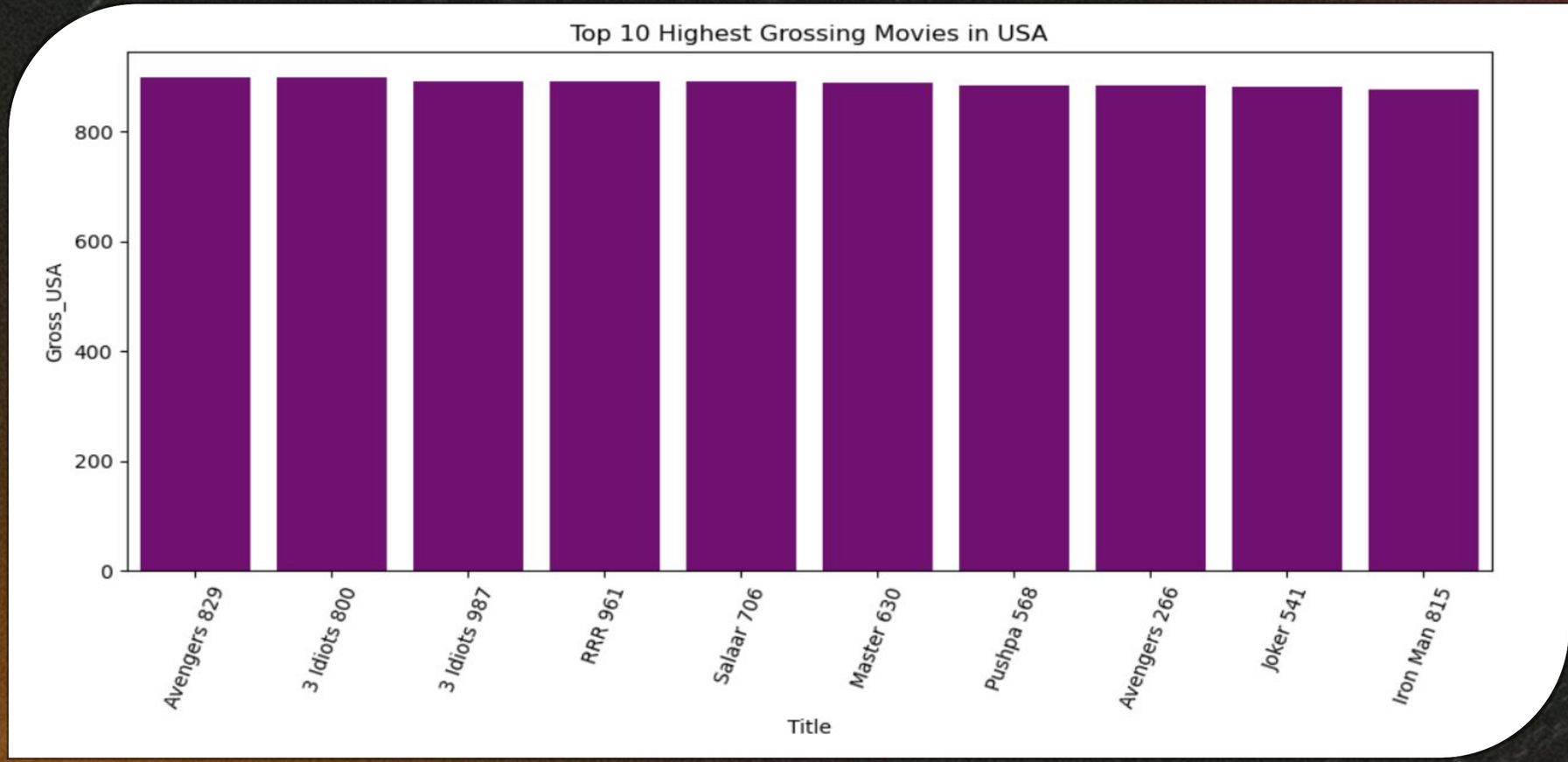
- The genre analysis shows **Action and Drama dominate the industry**.
- These genres have the highest number of movies, meaning they are the most preferred by both filmmakers and audience.
- Genres like Biography and Adventure have comparatively fewer releases.

Region Comparison



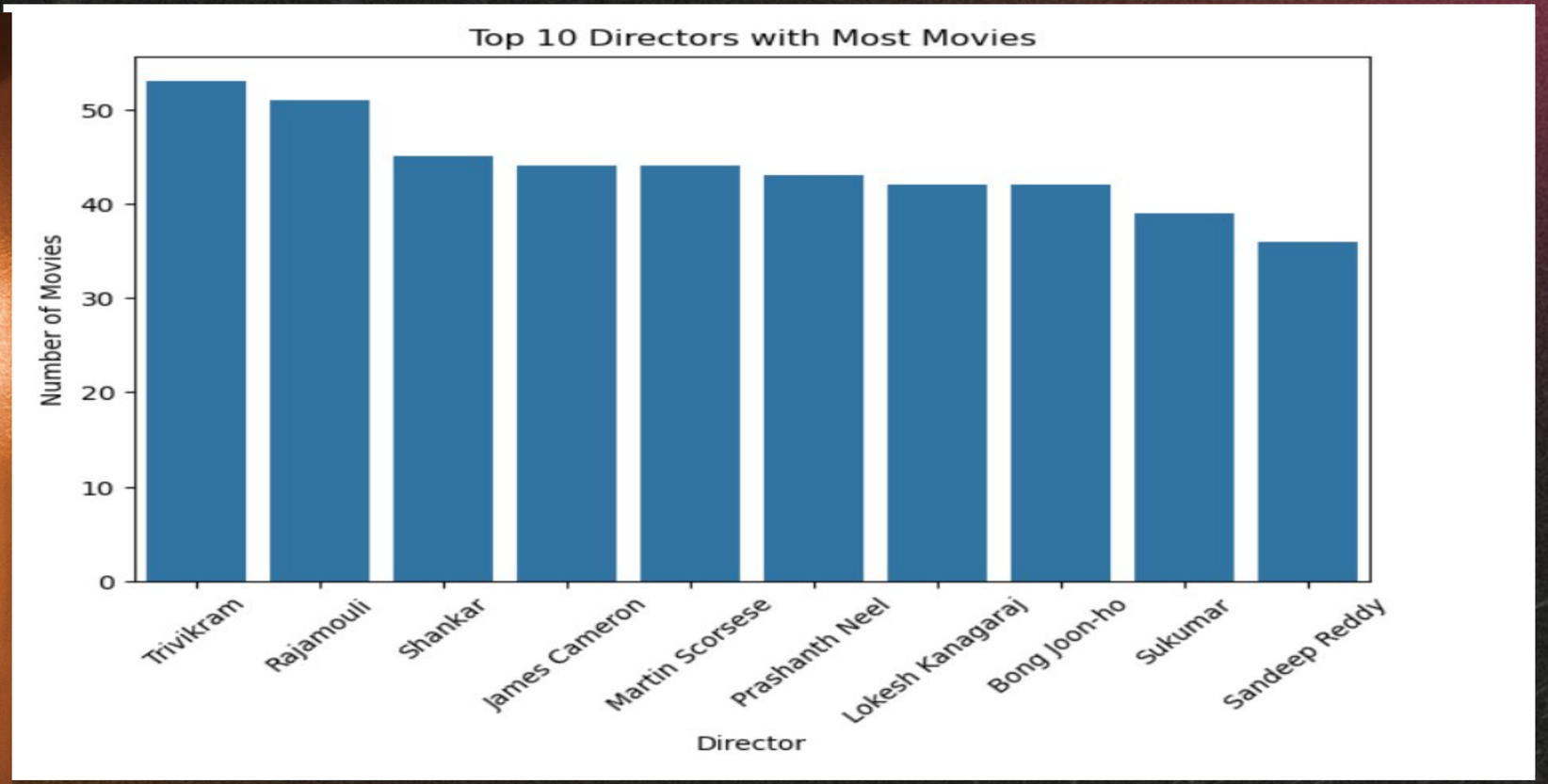
- From the region pie chart, **Hollywood contributes the highest number of movies**, followed by Tollywood and Bollywood.
- Hollywood dominates due to larger budgets, global distribution and advanced production value.

Top 10 Highest Grossing Movies



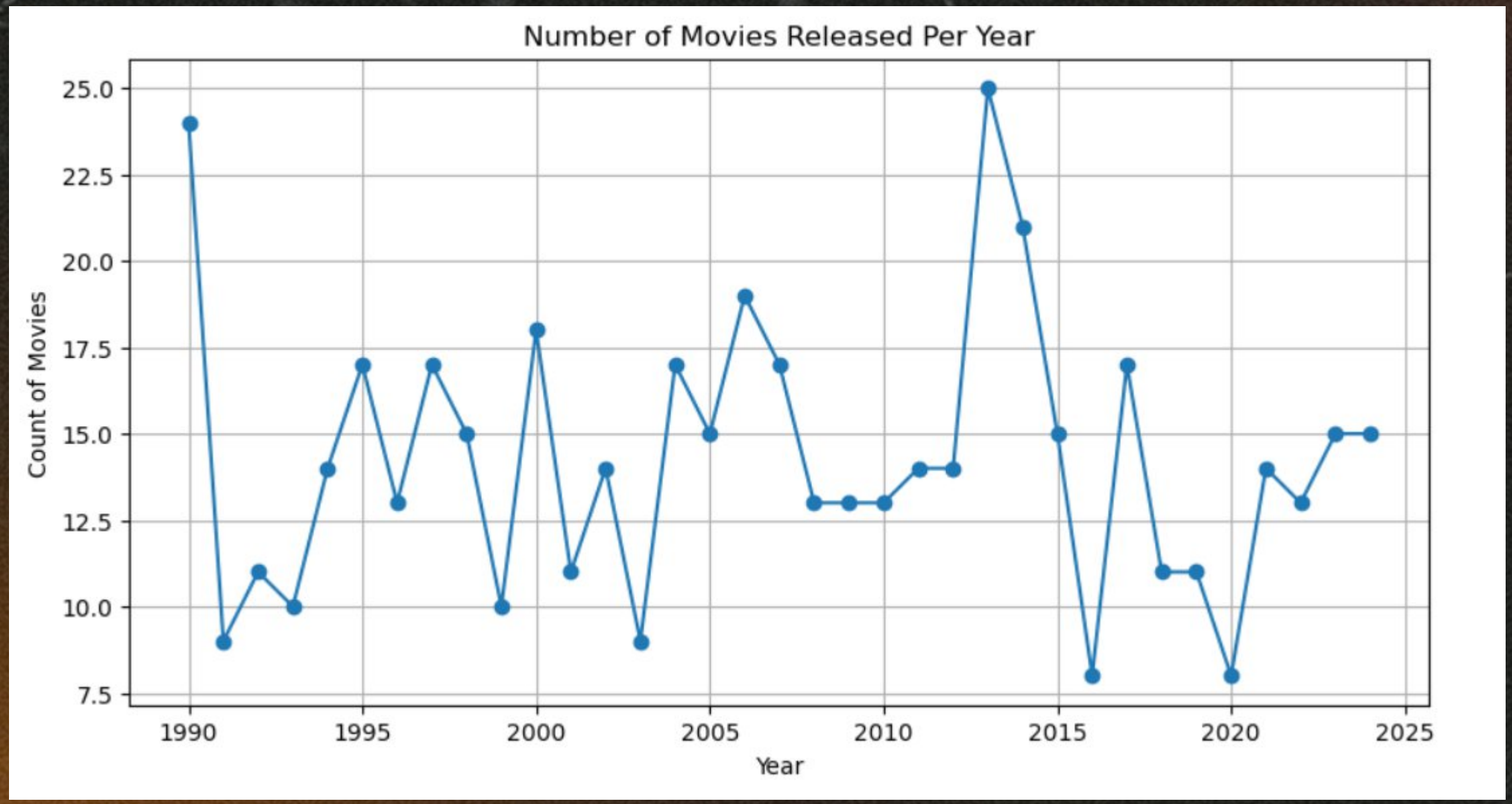
- The bar chart for Gross revenue shows that high-earning films mostly belong to **Hollywood** and **Action genre**.
- These movies earn more due to **franchise value, star power, marketing and worldwide release**.

Top 10 Directors with most movies



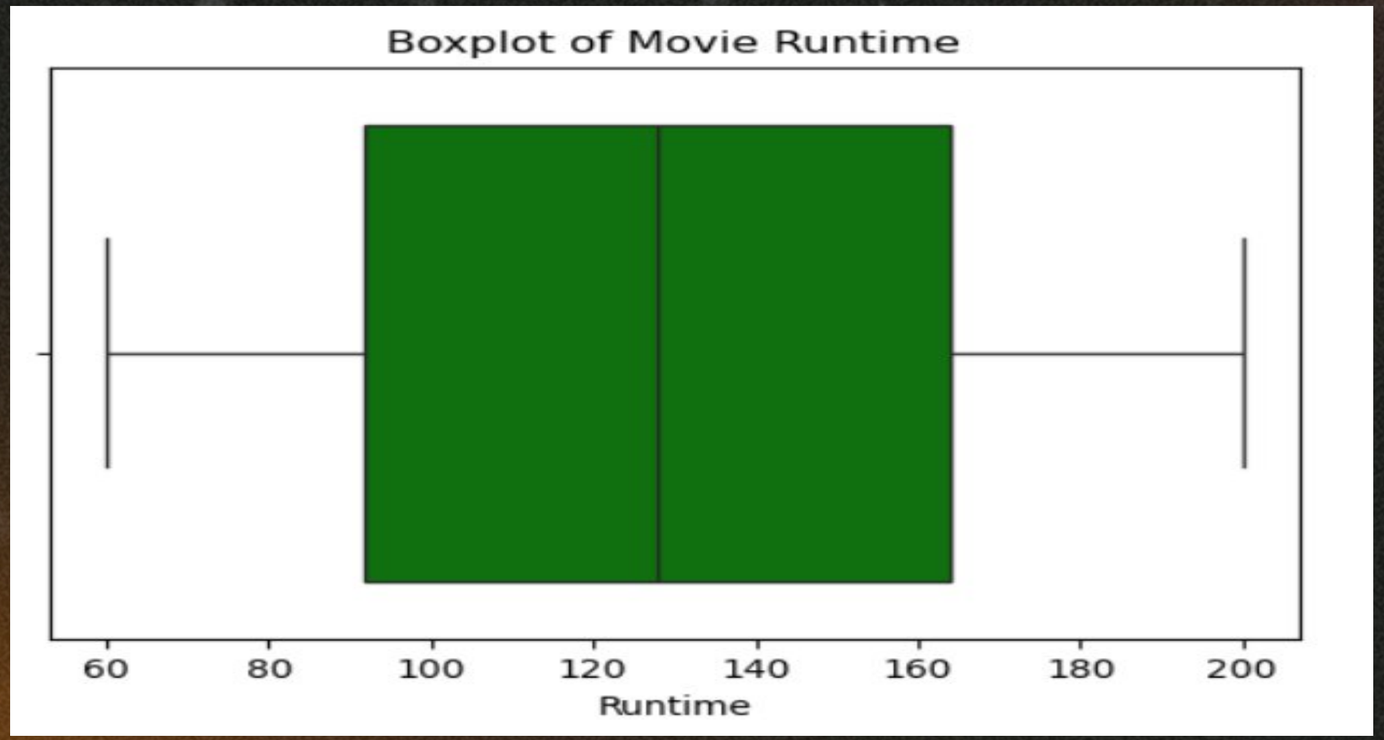
- The director analysis shows that directors with more films are typically **successful and consistent**, making them highly preferred by production houses.

Number of Movies Released Per Year



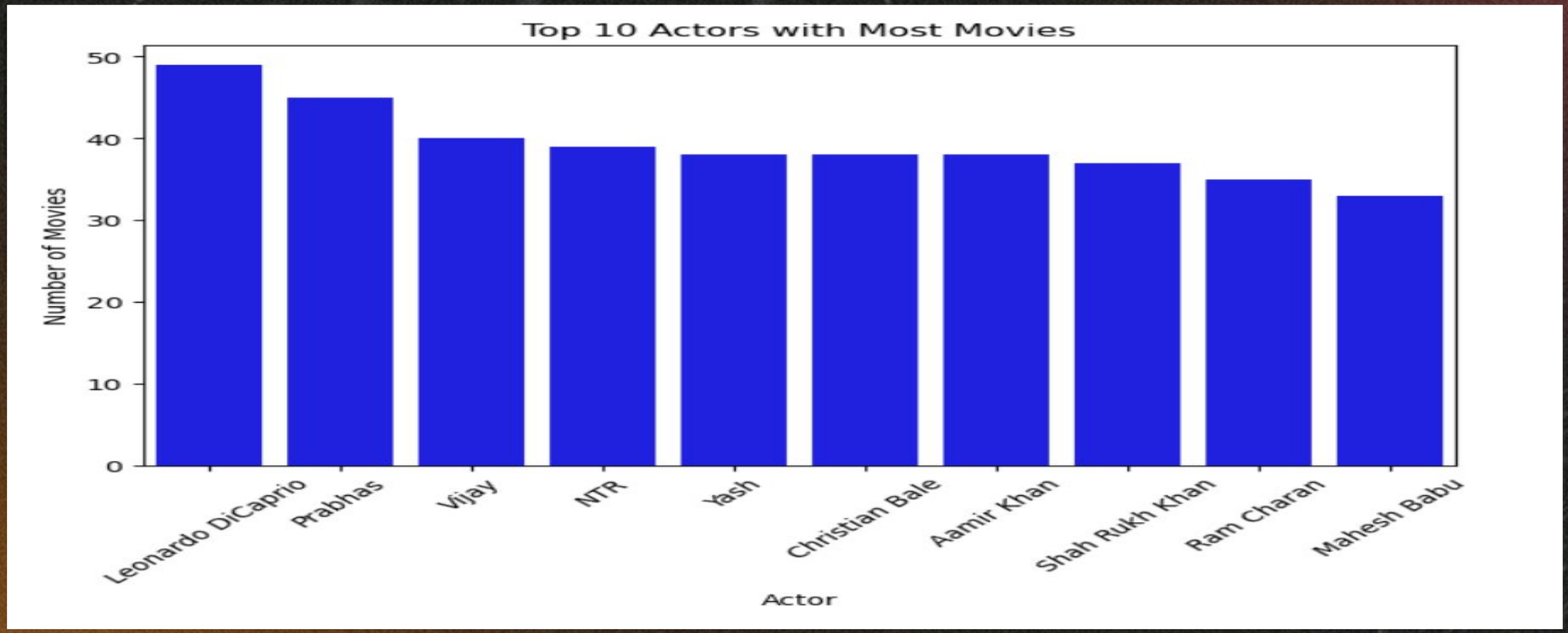
- As shown in the line plot, the number of films produced per year gradually increased over time.
- This highlights the **growth of the movie industry**, expansion of **OTT platforms**, and **audience demand** for content.

Movie Runtime Distribution



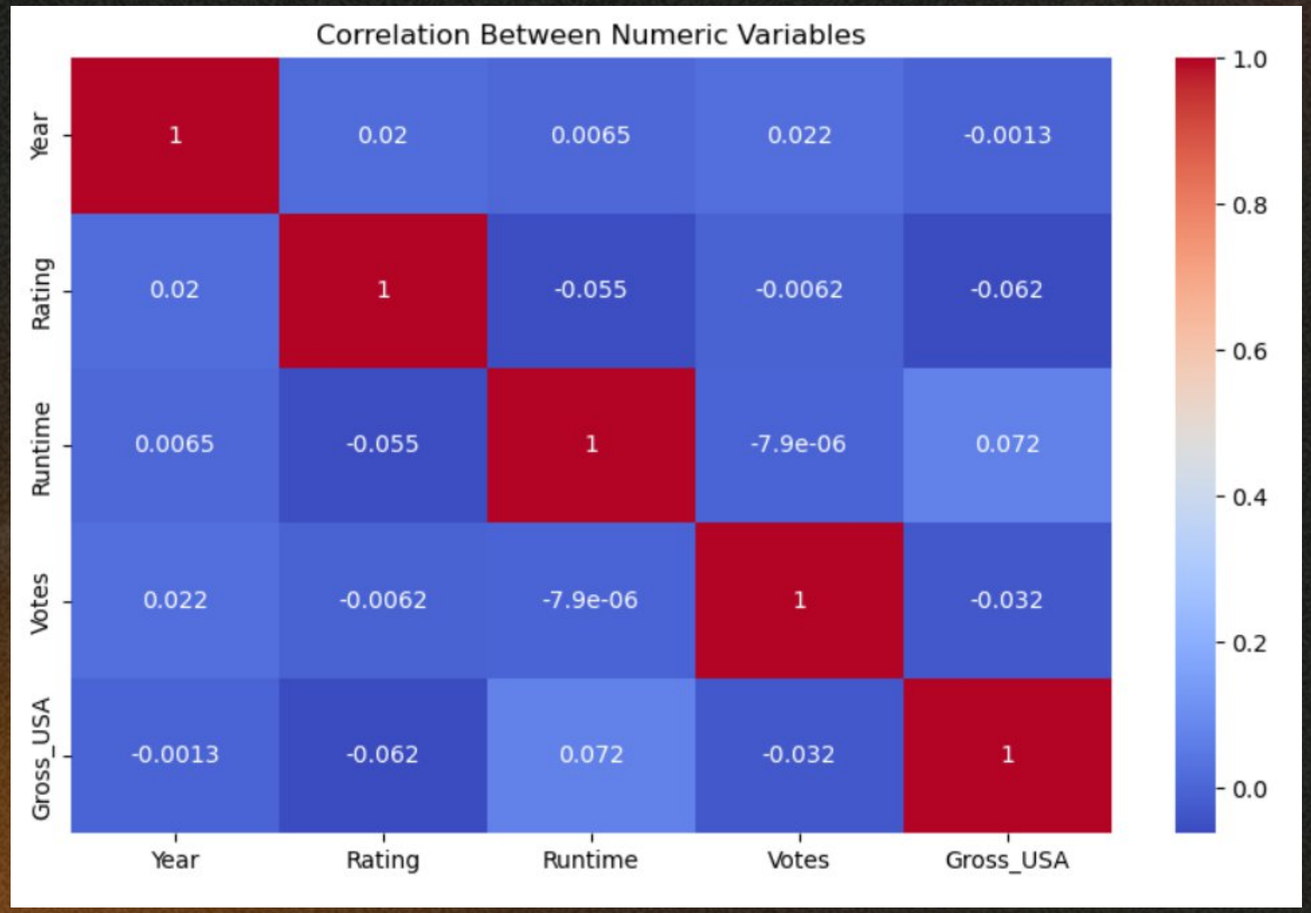
- The runtime distribution indicates most movies are between **120 and 160 minutes**.
- This suggests filmmakers follow a standard duration between **2 to 2.5 hours** to maintain audience engagement.

Top 10 Actors



- From the actor frequency chart, the top actors appear in multiple movies, proving their **high popularity and demand** in the industry.

Correlation Heatmap



- The heatmap indicates a strong positive correlation between **Votes and Gross revenue**.
- That means movies that receive higher audience attention tend to generate higher earnings.
- Runtime has weak correlation, meaning **longer movies do not guarantee success**.

SUMMARY OF OBSERVATIONS

- **Goal:** Understand movie ratings and trends.
- **Clean Data:** Removed unnecessary columns and fixed missing values.
- **Prepare Data:** Extracted numeric ratings and movie release years.
- **Analyze:** Checked rating patterns and relationships with genres.
- **Visualize:** Used boxplots and histograms to see ratings and outliers.
- **Insights:** Most movies have average ratings; some are very high or low; some genres rate higher than others.

CONCLUSION

Movie success mostly depends on genre, IMDB rating and audience reach (votes). Action and Drama movies perform well, Hollywood movies show higher revenue, and most movies have runtime between 120–160 minutes. More votes lead to higher earnings.

thank you

PRESENTED BY :

M.ANIL KUMAR

A.MALLA REDDY

INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS