**Movies Dataset Analysis Project Report**

**1.Introduction:**

This project analyzes a collection f movies released between the years 2006 and 2016. The objective is to understand how different factors such as Genre, Rating, Revenue, Votes, and Year influence movie performance.

**2. Dataset Overview:**

Data Used: IMDB DATASET.

Size: 1000 rows × 12 columns

* Columns :
* Title : Movie Name
* Genre : Movie Genres
* Description : Short Summary of the Movie
* Director : Director Name
* Actors : Cast Actors
* Year : Year of Release
* Runtime (Minutes) : Duration of the Movie
* Rating : IMDB Rating
* Votes : Number of Audience votes
* Revenue (In Millions) : Box office collection in Millions
* Metascore : Critic score
* Data Types :
* Continuous= ["Runtime (Minutes)", "Rating", "Revenue (Millions)", "Metascore"]
* Categorical= ["Title", "Genre", "Description", "Director", "Actors"]
* Count= ["Rank", "Votes", "Year"]

**3.** **Data Cleaning and Preprocessing**:

Before analyzing the dataset, a few cleaning steps were required to make the data accurate, consistent, and ready for visualization.

**a) Handling Missing Values**

* The columns **Revenue (Millions)** and **Metascore** contained missing values.
* **Revenue** is an important measure of movie performance, so removing rows would lead to data loss.  
  → Therefore, missing values in *Revenue* were **filled with the average revenue**.
* **Metascore** values were missing for some movies. Since Metascore differs a lot between movies, the missing values were **filled using the median score** to avoid misleading visuals

**b) Checking and Removing Duplicates**

* The dataset was checked for duplicate movie entries.
* No major duplicates were found, so no rows were removed.
* This ensured every movie appears only once in the dataset.

**c) Fixing Data Types**

* The dataset was reviewed to make sure every column had the correct data format.
* Numerical columns like **Votes**, **Rating**, and **Revenue** were kept in numeric format for calculations.
* This step ensured smooth grouping, comparison, and plotting during analysis.

**d) Outlier Treatment**

During the analysis, the **Rating** column was found to contain a few lower values. To ensure the data was consistent and not influenced by these values, we applied the **Interquartile Range (IQR) method**.

* **IQR = Q3 - Q1**
* **Lower Limit = Q1 - 1.5 × IQR**
* **Upper Limit = Q3 + 1.5 × IQR**

Any rating values **below the Lower Limit** or **above the Upper Limit** were treated as **outliers**.  
Instead of removing them, we **capped** the outliers:

* Values below the **Lower Limit** were replaced with the **Lower Limit**.
* Values above the **Upper Limit** were replaced with the **Upper Limit**.

This approach helped maintain the dataset size while ensuring that lower values do not affect the results.

**e) Creating Rating Groups**

* To make rating comparisons easier, the continuous IMDB ratings were grouped into categories.
* A new column **Rating\_Group** was created using the following scale:

| **IMDb Rating** | **Rating Group Name** |
| --- | --- |
| 0.0 – 4.9 | **4 Star** |
| 5.0 – 5.9 | **5 Star** |
| 6.0 – 6.9 | **6 Star** |
| 7.0 – 7.9 | **7 Star** |
| 8.0 and above | **8 Star** |

* This helped in analyzing how movies with similar rating levels perform across years and genres.

**f) Final Prepared Dataset**

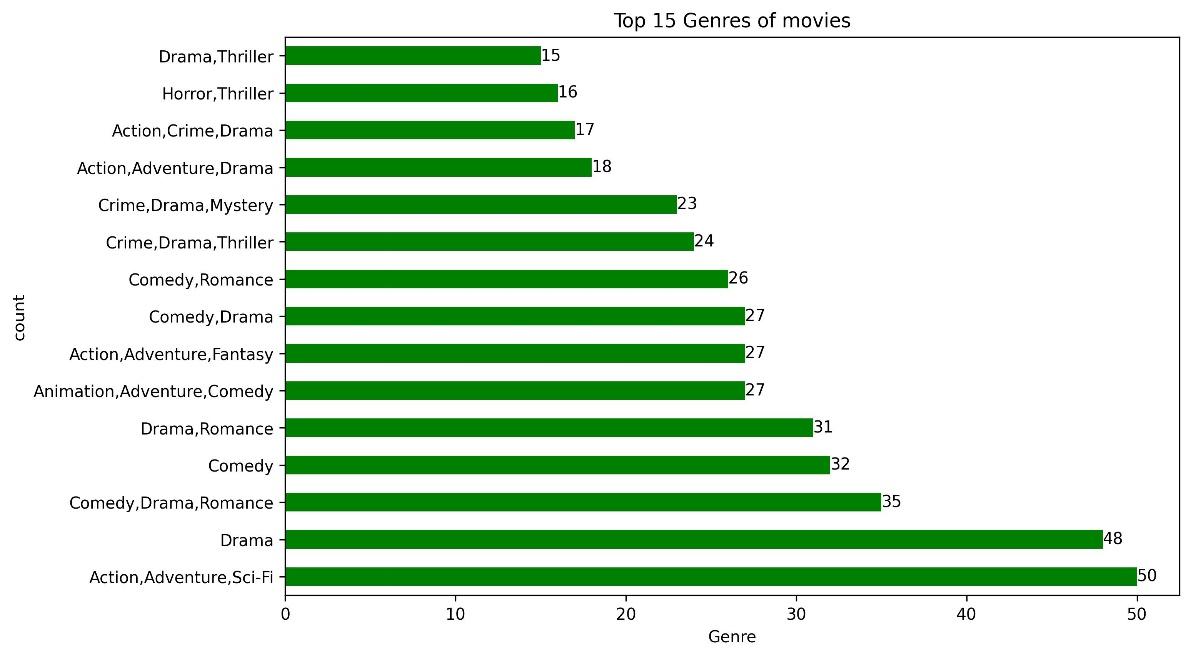
After completing the above cleaning steps:

* There were no missing important values,
* Data formats were correct,
* Treated Outliers,
* And ratings were grouped.

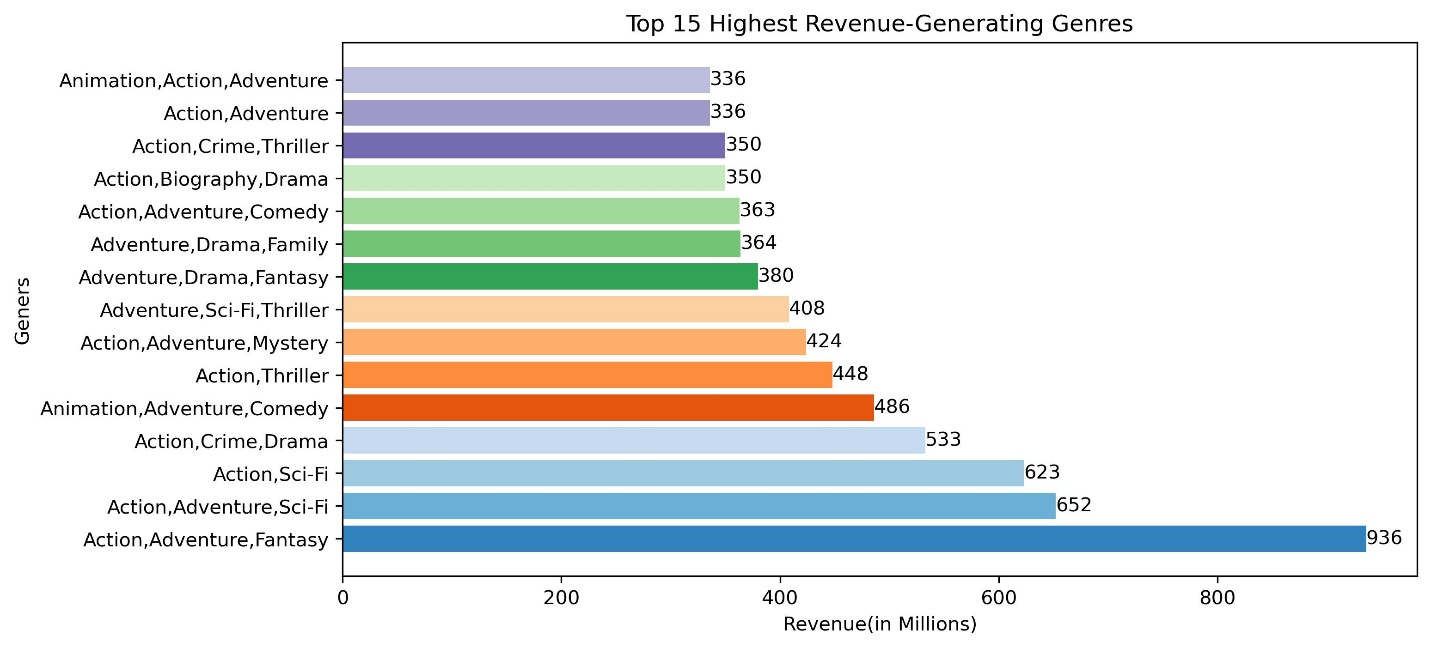
The dataset became **clean, structured, and ready for detailed Exploratory Data Analysis (EDA).**

**4. Observations and Key Insights:**

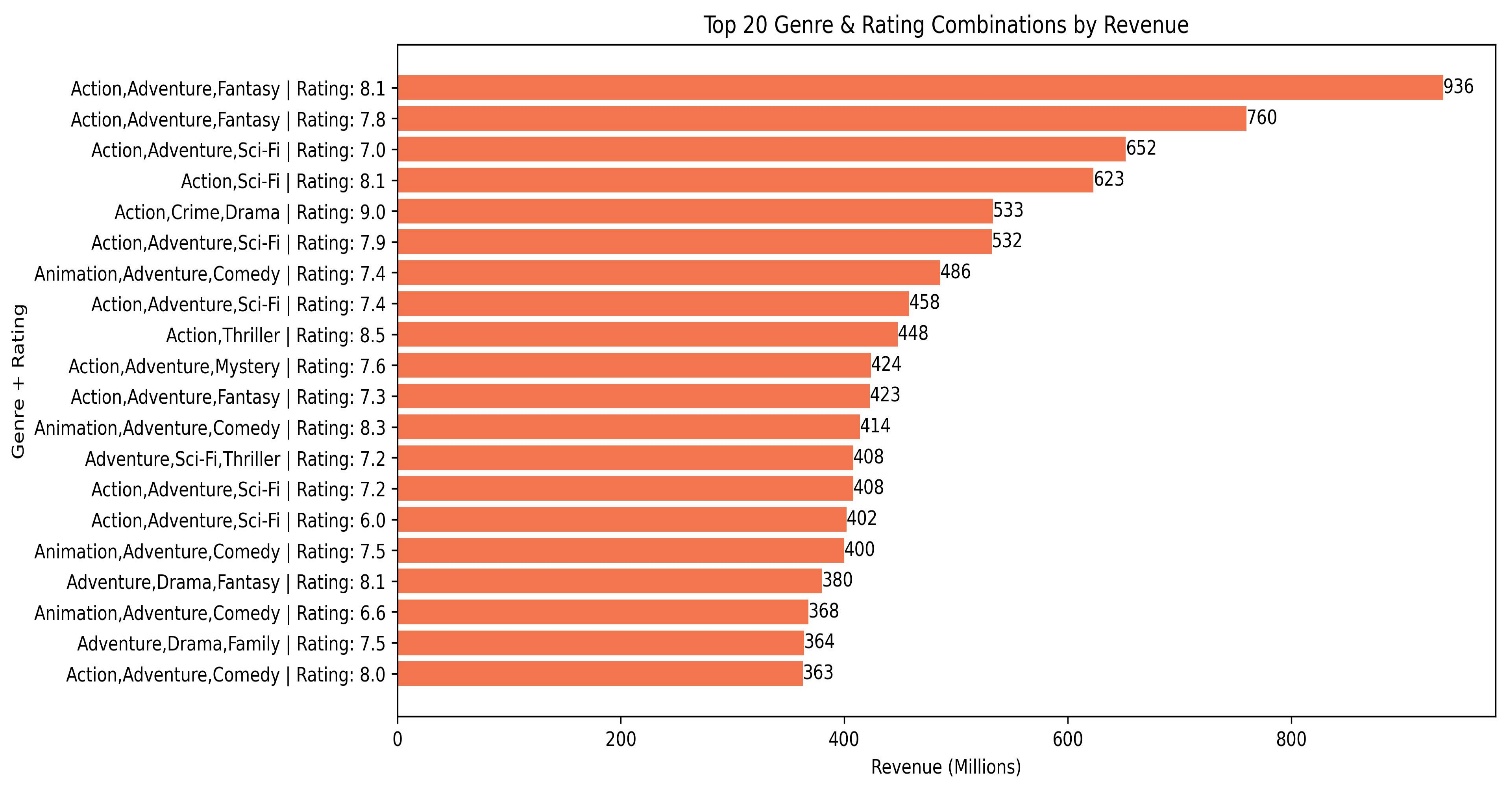
* **1. Genre wise Analysis**.
* **Genre wise distribution**



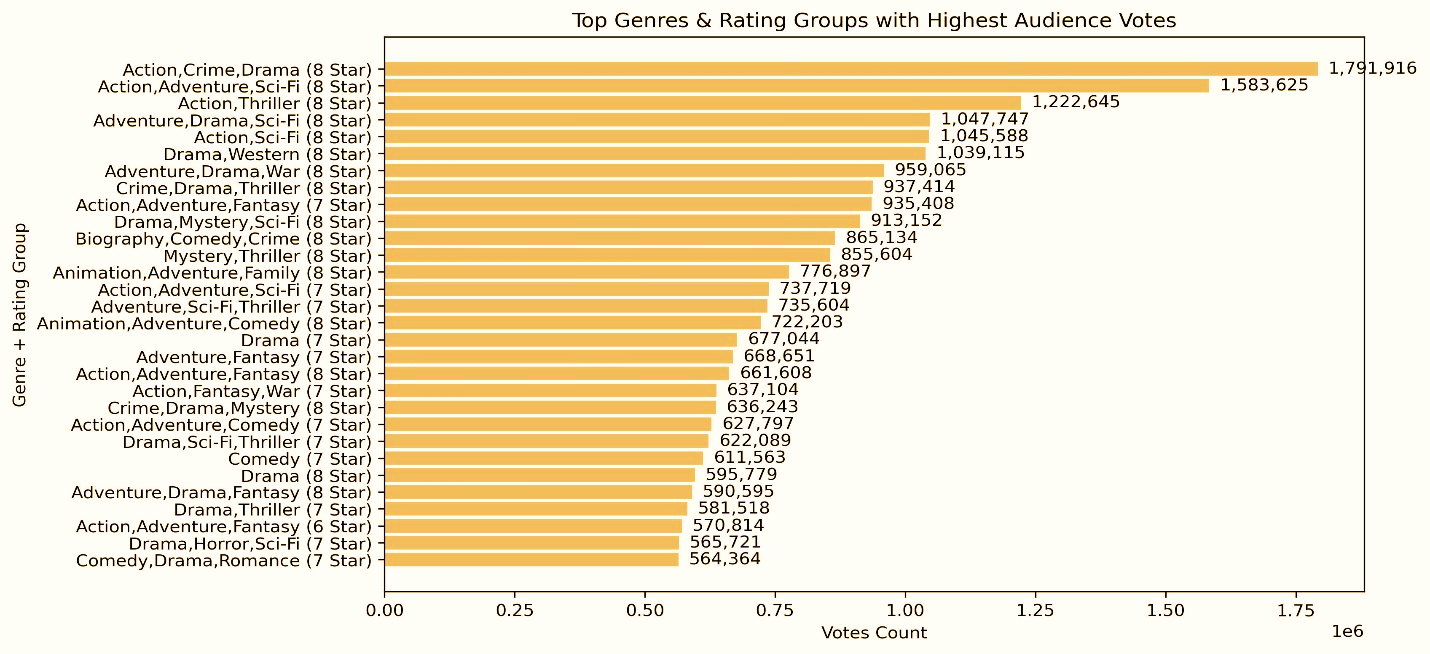
* Genre wise Revenue



* Genre with rating and revenue generating



* Genre wise voting and Rating



* **Insight**:

1. **Action–Adventure–Sci-Fi and Fantasy** movies dominate across production, revenue, and audience engagement, showing that viewers prefer large-scale, visually rich, and franchise-based films.

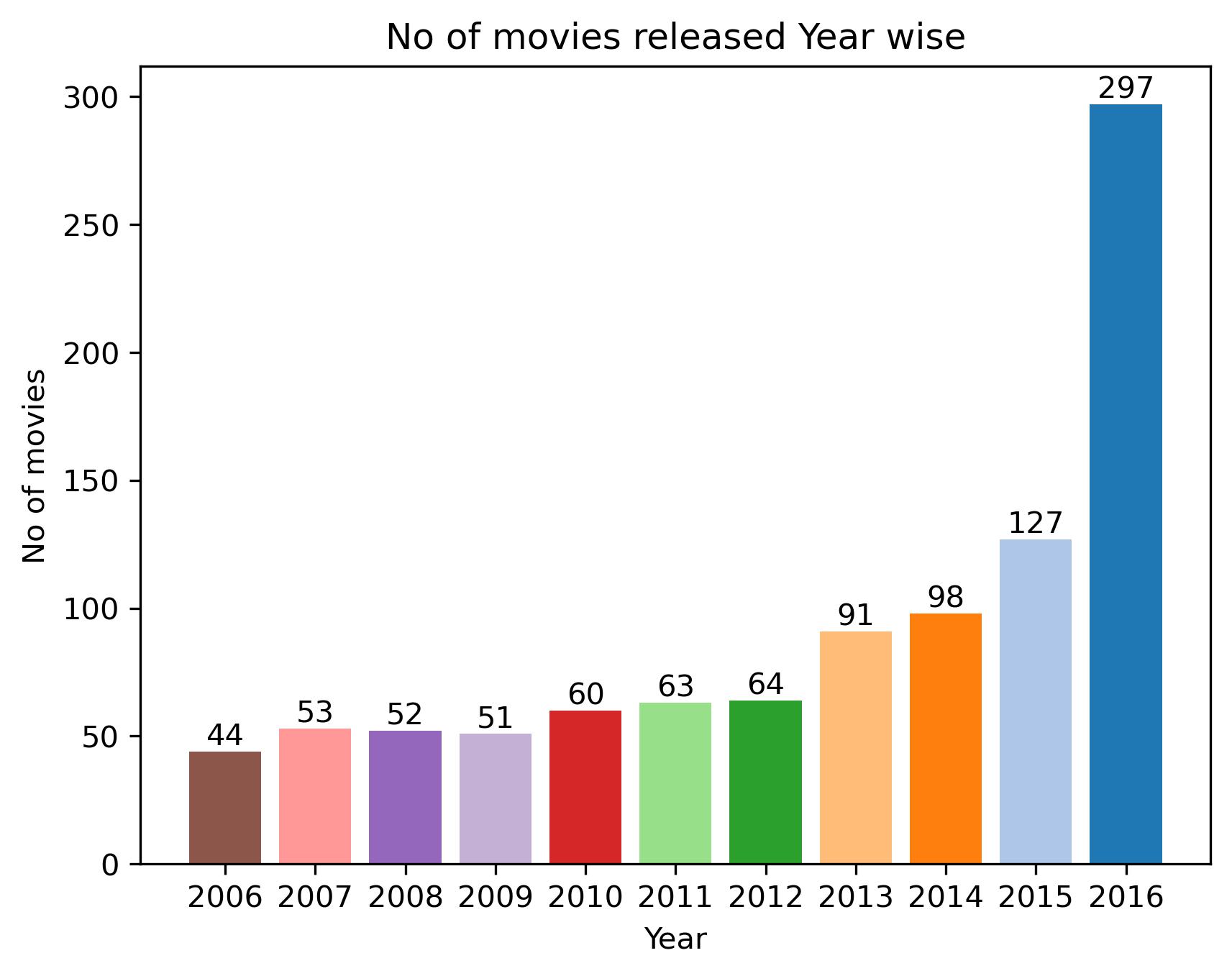
2. **Drama** and **Comedy–Romance** genres remain popular, proving that emotional and story-driven movies continue to connect with audiences.

3. **Action–Adventure–Fantasy** earned the **highest revenue (~$936M)**, while **Animated Adventure–Comedy** films also perform well, especially among family audiences.

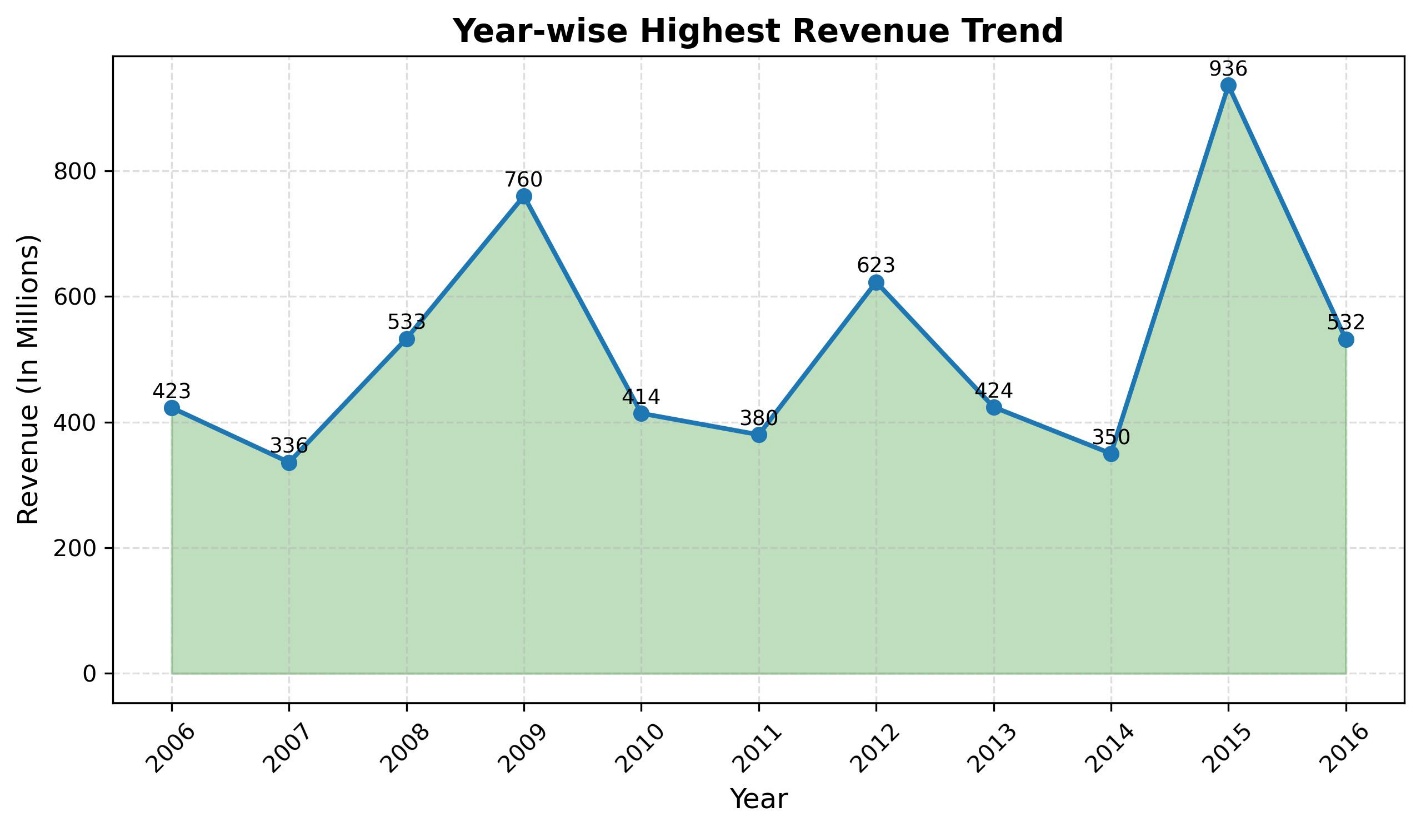
4. Box-office success is influenced more by **franchise power and cinematic appeal** than IMDB ratings, as even mid-rated films achieve huge earnings.

5. Overall, audiences are drawn to two main types of movies — **spectacle-driven blockbusters** and **emotionally engaging dramas or comedies**, both contributing strongly to the film industry’s success.

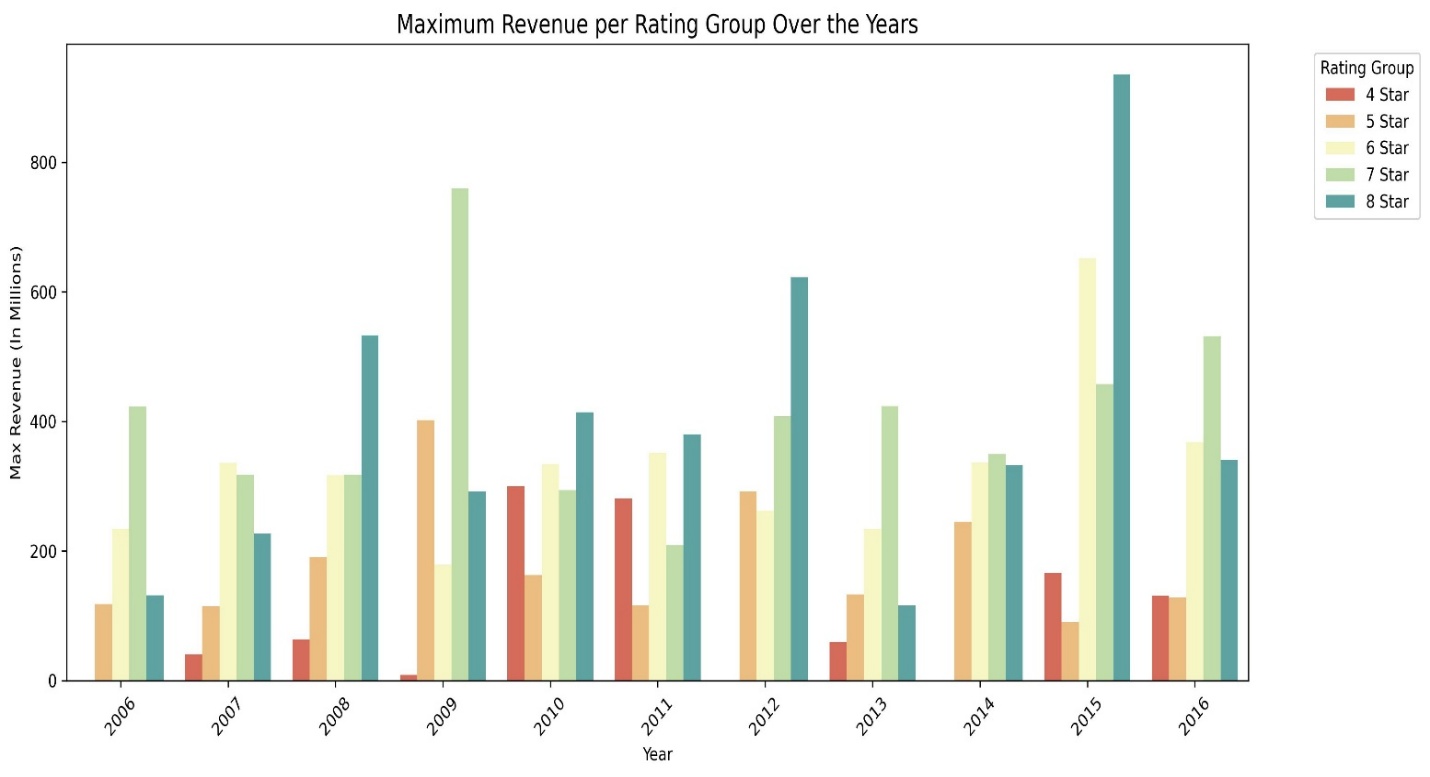
* **2. Year Wise Analysis**:
* Year wise movie count.



* Yearly Revenue Trend Generated by movies



* Year wise movie rating along with revenue.



* **Insight**:

1. Between **2006 and 2016**, the movie industry saw a significant **increase in production**, with the number of movies released in 2016 being about **6–7 times higher** than in 2006.

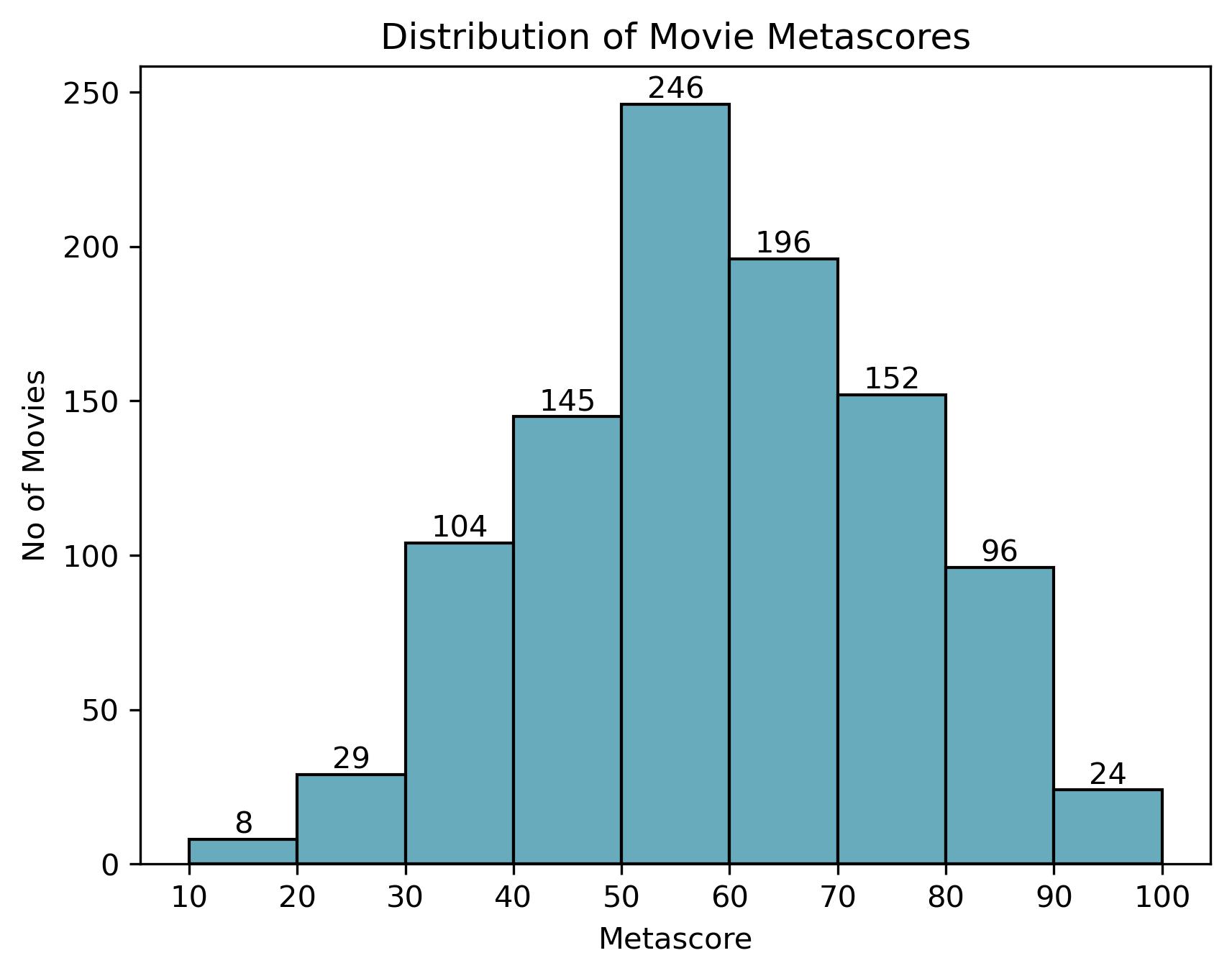
2. Movies with **higher IMDb ratings (7★ and 8★)** consistently earned the **highest revenues**, proving that audience quality perception directly impacts box-office success.

3. **Revenue trends** fluctuated over the decade, with notable peaks in **2009, 2012, and 2015**—indicating blockbuster releases in those years, especially in the higher rating groups.

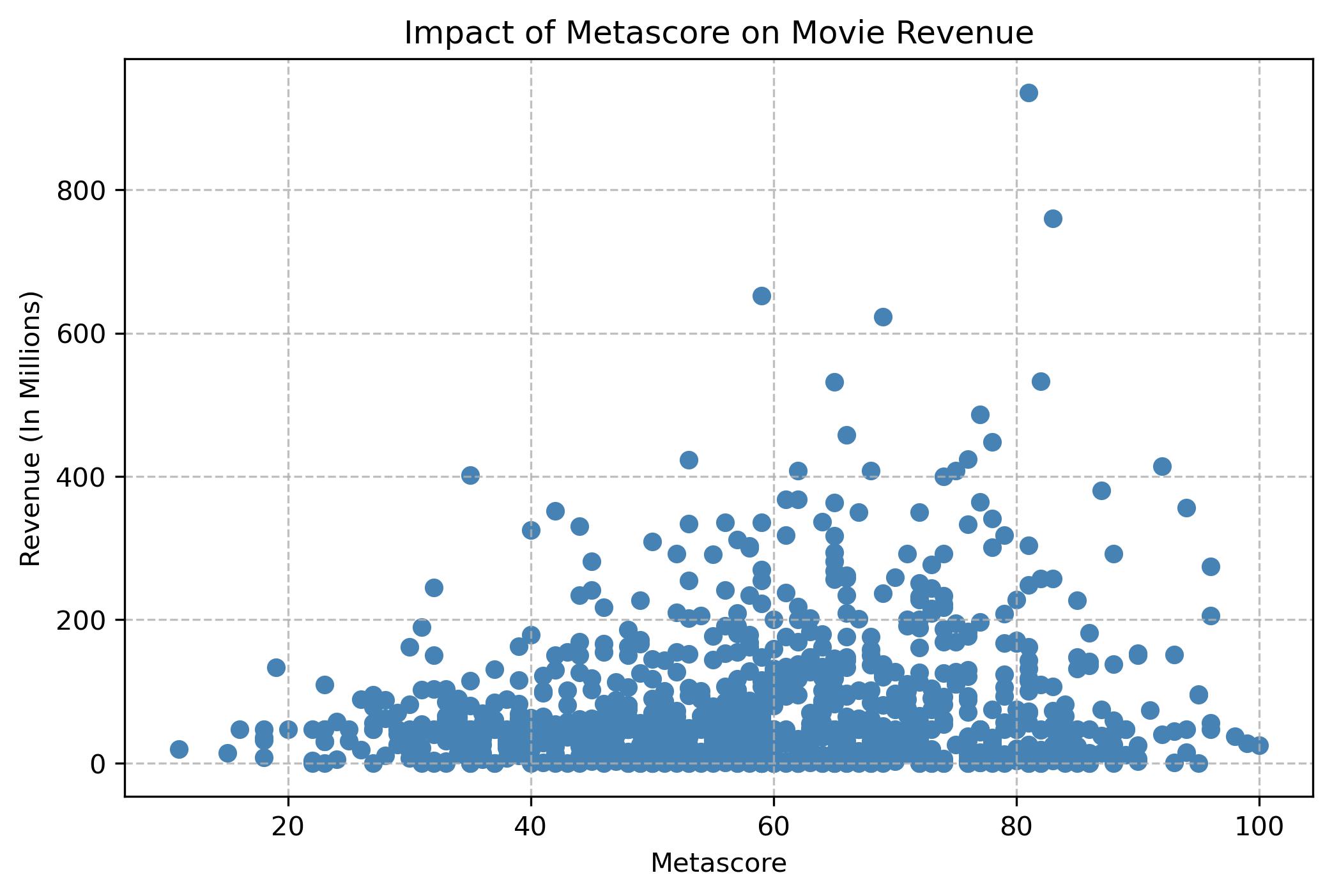
4. The **year 2015 recorded the highest revenue (~$936 million)**, showing a major box-office boom and possibly the influence of big franchise or sequel releases.

5. Overall, the data highlights a clear **positive relationship** between **higher-rated movies**, **increased production**, and **greater revenue potential**— . showing that the film industry expanded in both production scale and audience preference for quality content.

* **3. Metascore wise revenue:**
* Frequency of Metascores of the movies.

****

* Metascore with revenue

****

**Insight:**

1. The majority of movies have **Metascores between 50 and 60**, indicating that most films receive **moderately positive reviews** from critics.

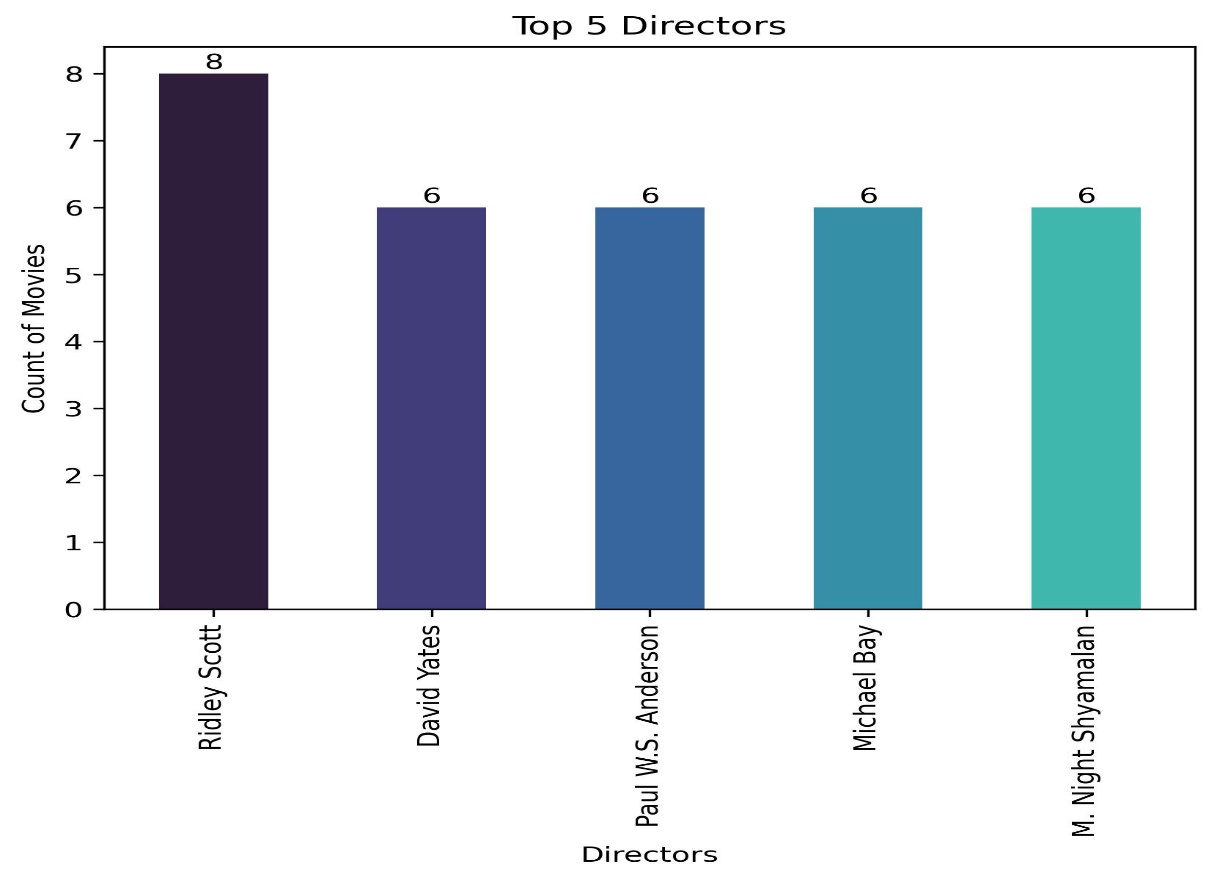
2. Movies with **very low (<30)** or **very high (>80)** Metascores are comparatively rare, suggesting that extreme critical opinions are uncommon.

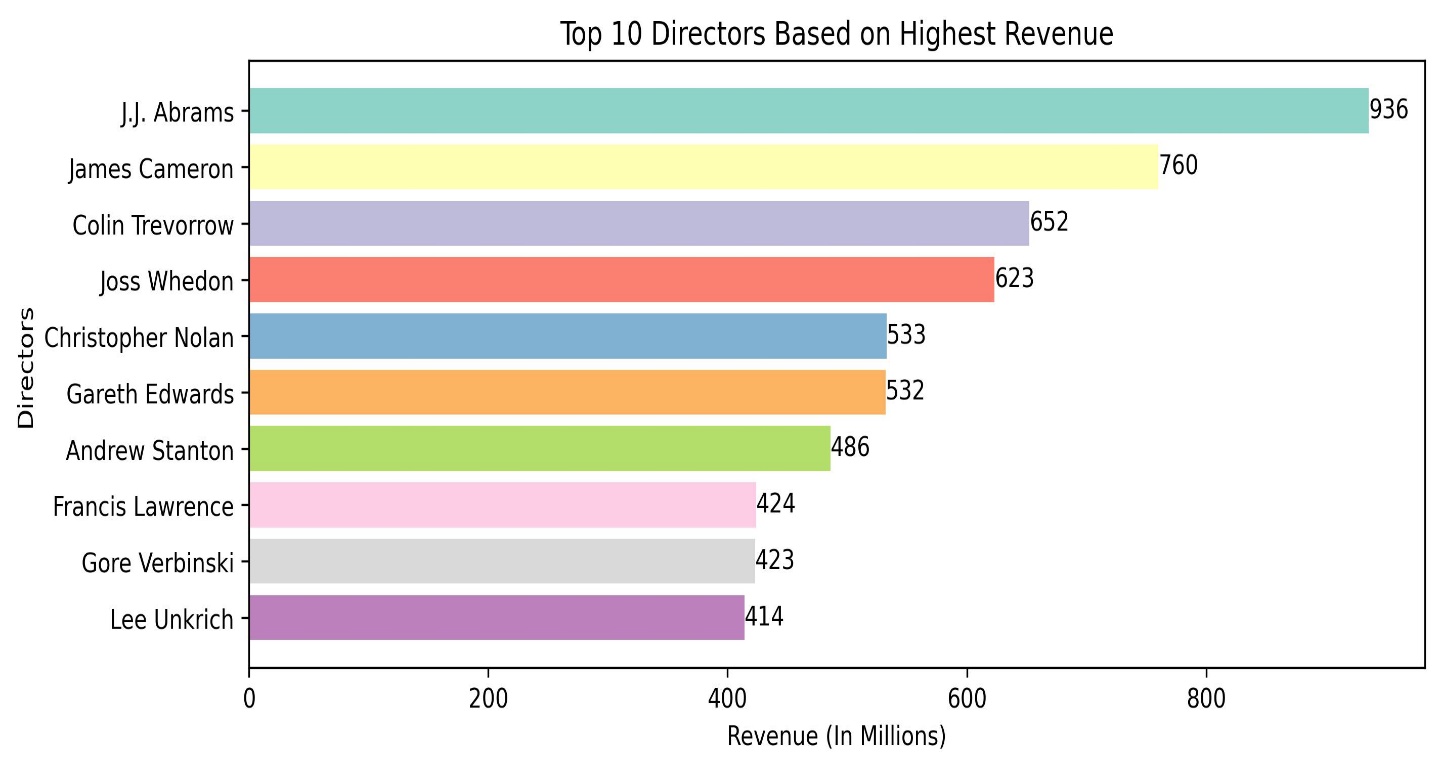
3. The **average Metascore (~58)** reflects that most movies are viewed as fair or average in quality by critics.

4. The **Metascore vs Revenue** comparison reveals **no strong correlation** between critical ratings and box-office performance.

5. This pattern suggests that **commercial success is influenced more by audience engagement, marketing strength, and franchise popularity** rather than critic reviews.

**4.Top Directors:**

****

****

* **Insight:**

1. **Ridley Scott** directed the most movies (8), followed by **David Yates**, **Paul W.S. Anderson**, **Michael Bay**, and **M. Night Shyamalan** with 6 each.

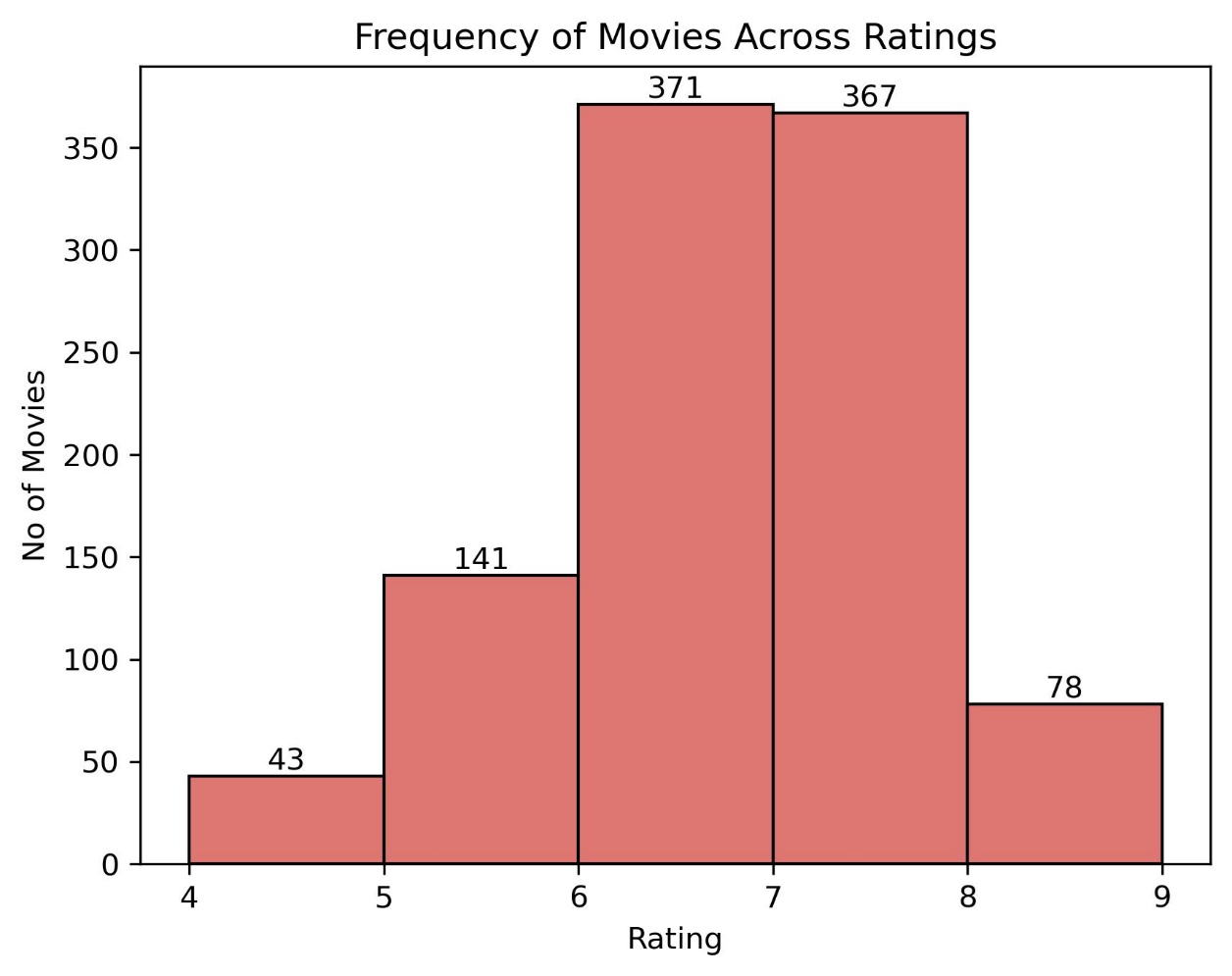
2. Only a few directors consistently produce multiple films, showing strong industry presence and productivity.

3. **J.J. Abrams** achieved the **highest revenue (~$936M)**, followed by **James Cameron (~$760M)** and **Colin Trevorrow (~$652M)**.

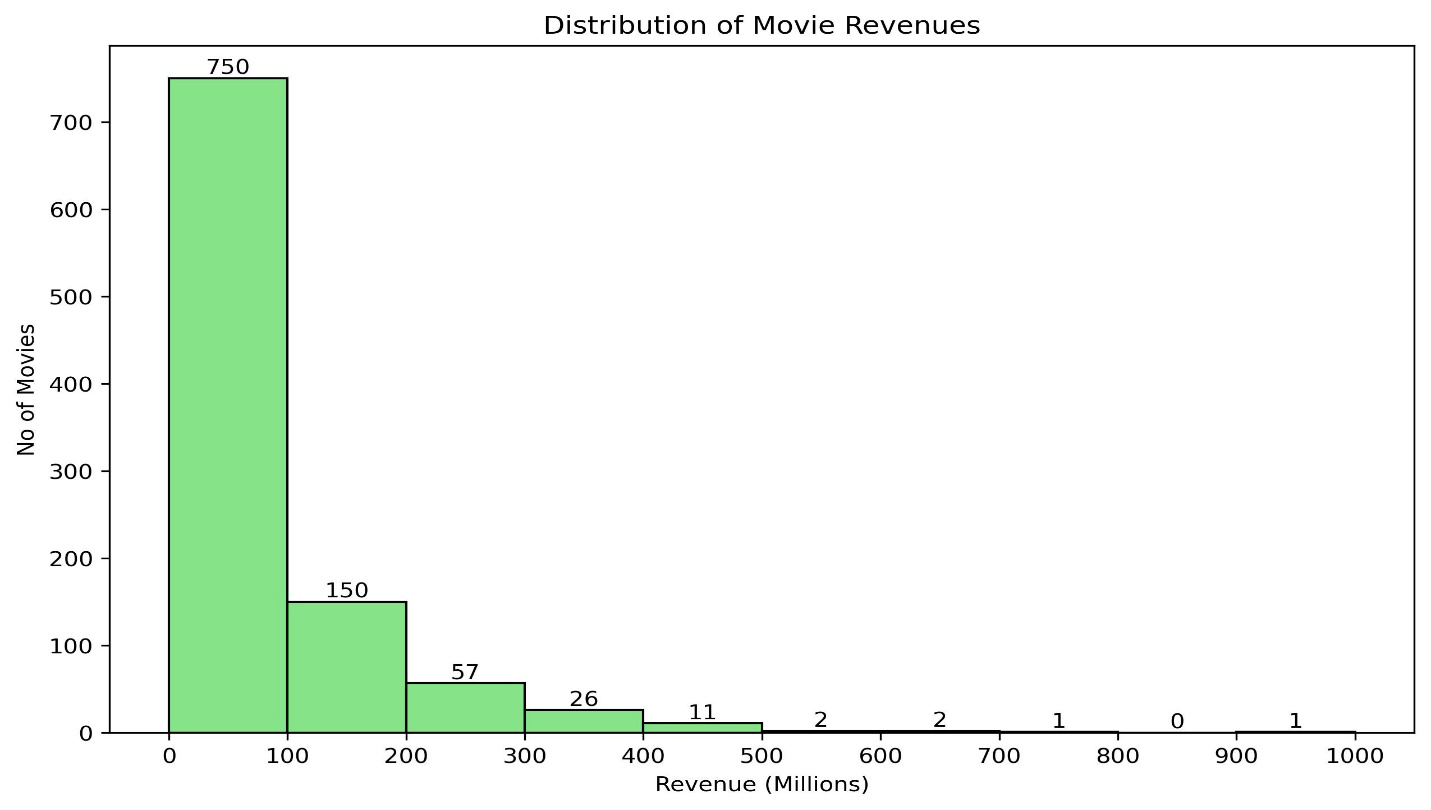
4. Directors with large-scale and visually engaging movies, like **Christopher Nolan** and **Joss Whedon**, also generate high box-office results.

5. Overall, **experienced and franchise-based directors drive both production volume and revenue success** in the movie industry.

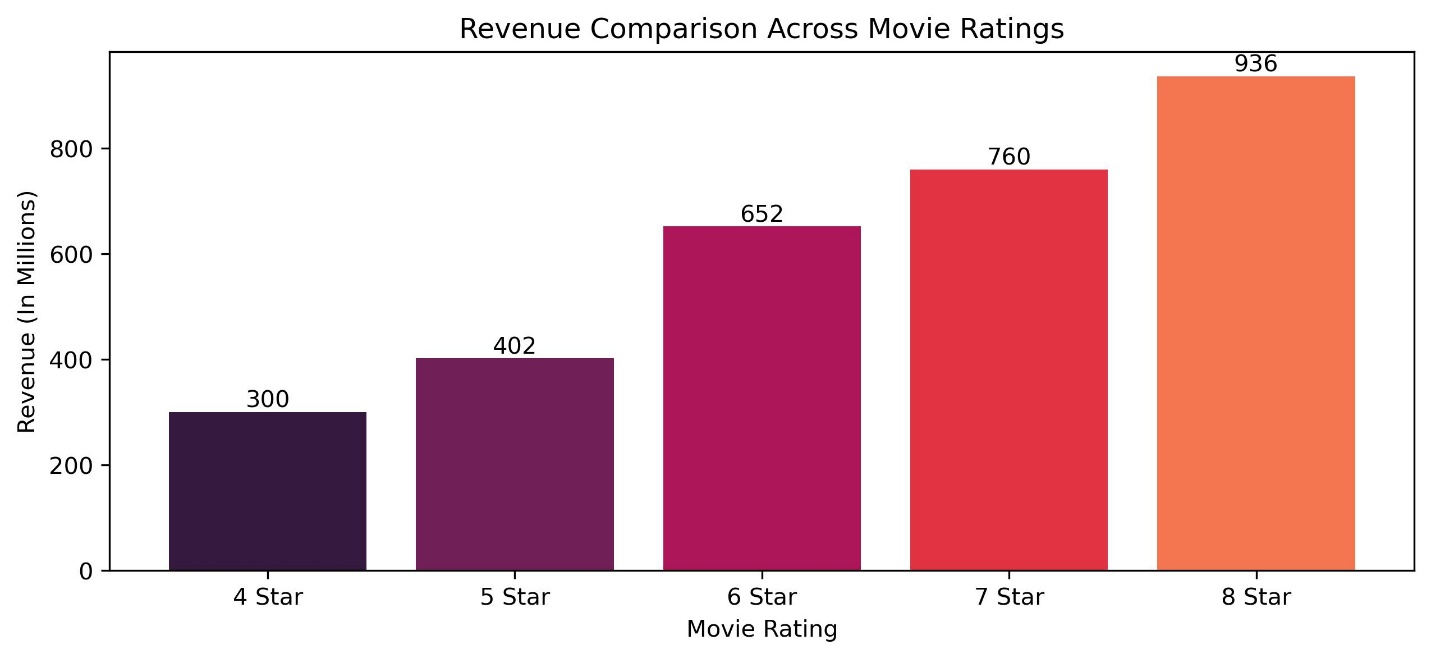
* **5. Revenue and Rating and votes:**
* Rating across all movies.

****

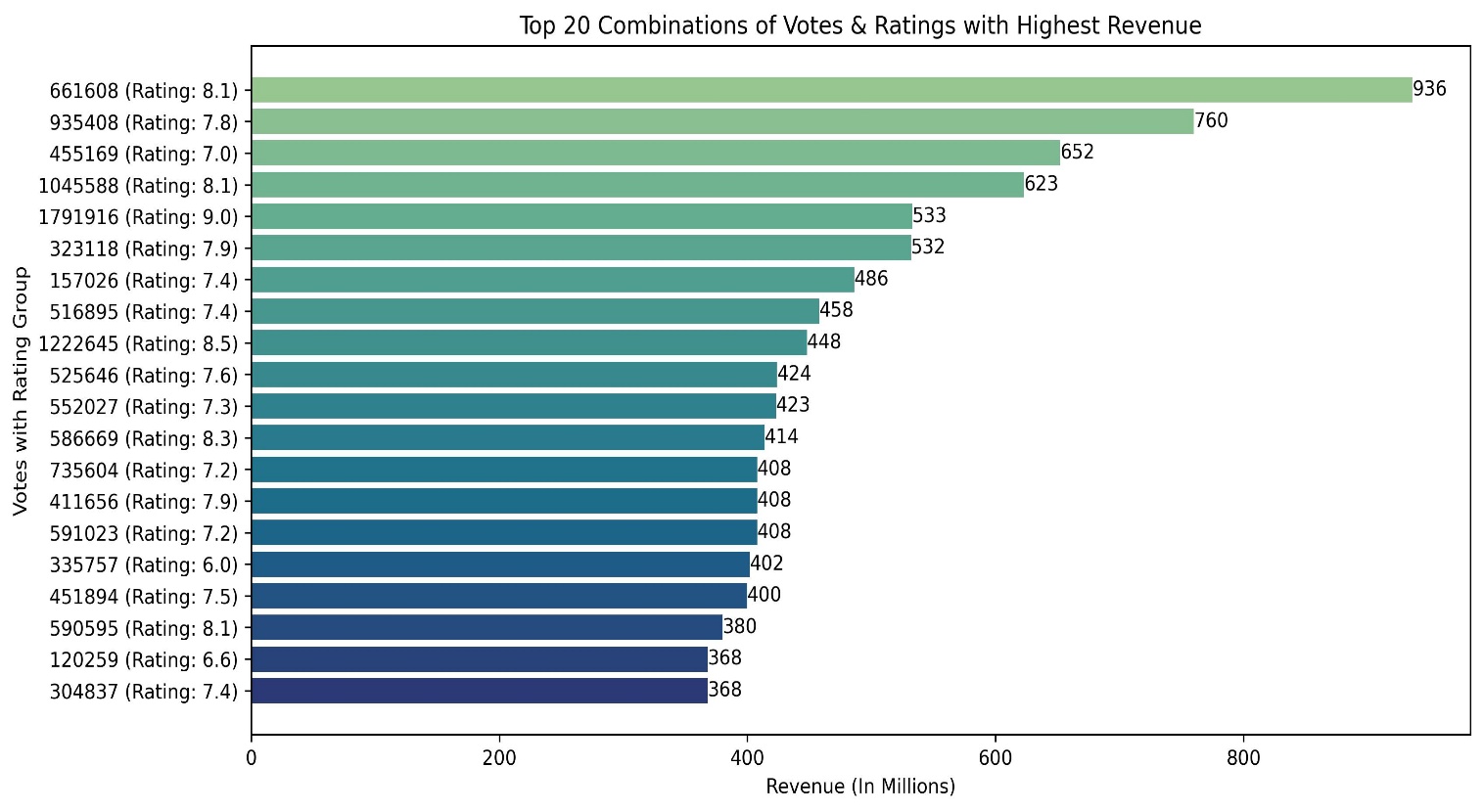
* Revenue collection of all movies.

****

* Rating wise revenue collection.



* Revenue with Votes Rating



* **Insight:**

1. **Movie Quality Distribution:**  
Most movies are rated between **6 and 8 stars**, meaning the majority of films are perceived as **average to good quality** by audiences. Very few movies fall below 6 (poor) or above 8 (excellent).

2. **Box Office Performance:**  
The majority of movies earn **below $100 million**, showing that most films achieve **moderate commercial success**. Only a few movies surpass the $200M–$500M range, and a single standout film reaches around **$936 million**, marking it as a blockbuster movie.

3. **Rating–Revenue Relationship:**  
There is a **clear positive correlation between ratings and revenue** — as ratings increase, revenue rises.

* 4★ movies earn around **$300M**
* 5★ → **$402M**
* 6★ → **$652M**
* 7★ → **$760M**
* 8★ → **$936M**  
  This suggests that **higher-rated movies attract larger audiences** and perform better financially.

4. **Audience Engagement (Votes) Impact:**  
Movies with **higher votes and ratings above 7.5** tend to generate **significantly higher revenue**.  
This indicates that when audiences enjoy a movie, they talk about it more, which helps increase its box-office earnings

5. **Overall Conclusion:**  
The analysis shows that **ratings and audience engagement play a vital role** in a movie’s financial success.

* **Quality storytelling (higher ratings)** and **strong audience reach (more votes)** together lead to blockbuster revenues.
* Most movies perform moderately, but those that achieve both **high quality and audience approval** stand out as major box-office hits.

* **Conclusion:**

This analysis of movies from **2006 to 2016** shows that audience taste and popularity strongly shape a movie’s success.  
Genres like **Action, Adventure, Sci-Fi, and Fantasy** earn the highest revenues because people enjoy big, exciting, and visually rich stories.  
Movies with **higher ratings (7★–8★)** usually make more money and get more votes, showing that good-quality films attract larger audiences.  
However, some lower-rated movies also do well because of **famous directors, strong marketing, or popular franchises**.  
Critic scores (Metascore) do not always match revenue — audiences, not critics, decide what becomes a hit.  
Famous directors such as **J.J. Abrams, James Cameron, and Ridley Scott** lead many top-performing movies, proving that experienced filmmakers often bring box-office success.  
Overall, movies that combin **quality storytelling, audience appeal, and franchise strength** achieve the greatest box-office success.