# **Exploratory Data Analysis Report: IMDb Top Movies Dataset**

## **1. Introduction**

This report presents an **Exploratory Data Analysis (EDA)** of the **IMDb Top Movies Dataset** (imdb\_movies\_clean\_full\_dataframe.csv). The primary goal of this EDA is to understand the structure of the data, uncover underlying patterns, identify anomalies, and derive insights related to movie ratings, runtime, and temporal trends. The analysis is based on the top-rated movies as compiled by IMDb.

## **2. Dataset Overview**

The dataset contains records of top-rated movies, characterized by the following columns:

| **Column Name** | **Description** | **Data Type** |
| --- | --- | --- |
| **rank** | The movie's rank on the IMDb list. | Integer |
| **title** | The title of the movie. | String |
| **year** | The release year of the movie. | Integer |
| **runtime\_min** | The duration of the movie in minutes. | Float |
| **certificate** | The movie's age rating/classification. | String |
| **rating** | The IMDb rating of the movie (out of 10). | Float |
| **votes** | The number of user votes contributing to the rating. | Float |
| **movie\_url** | The IMDb URL for the movie. | String |
| **decade** | The calculated decade of release. | Integer |

The dataset comprises **250 records** and is considered clean for this analysis, with no immediate missing values or data type inconsistencies noted for the core analytical variables (**rating, runtime\_min, year, decade**).

## **3. Data Distribution and Characteristics**

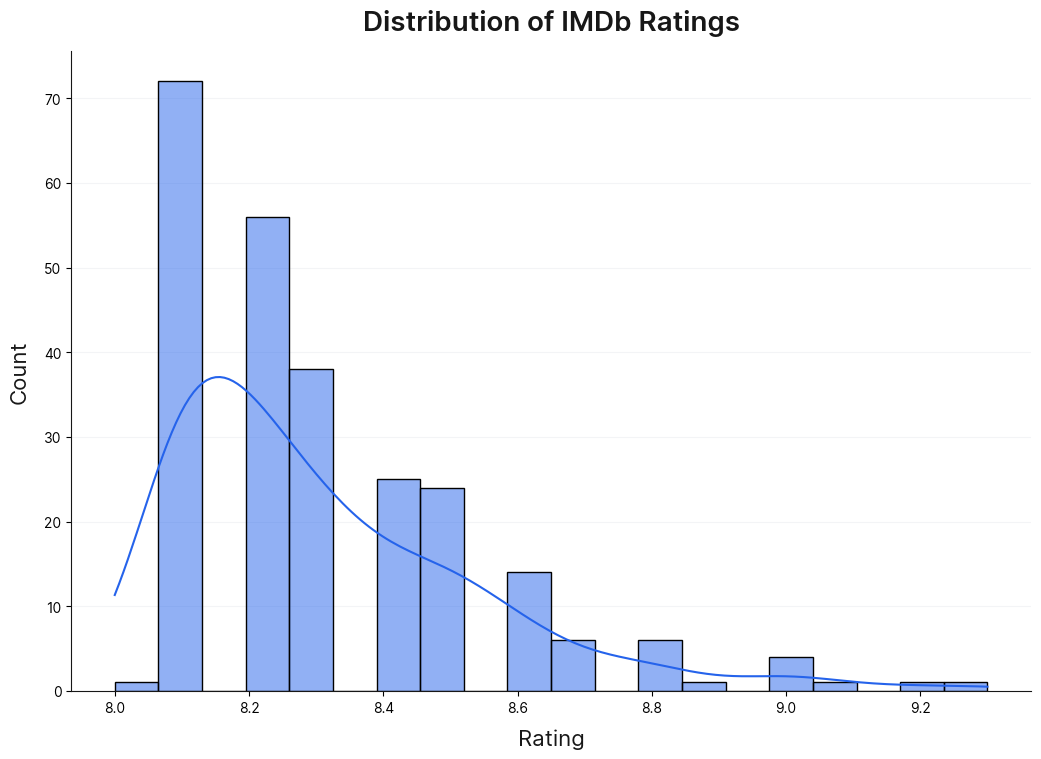
### **3.1. Distribution of Ratings**

The histogram below illustrates the distribution of IMDb ratings for the movies in the dataset.

* **Observation:** The distribution of ratings is **heavily skewed to the left** (negatively skewed), meaning the majority of the top 250 movies have very high ratings, clustered between **8.1 and 9.3**. This is expected, as the dataset is pre-filtered to include only the highest-rated movies. The peak frequency occurs in the range of 8.1 to 8.4.

### **3.2. Top 10 Movies by Rating**

The bar chart highlights the top 10 movies in the dataset, ordered by their IMDb rating.

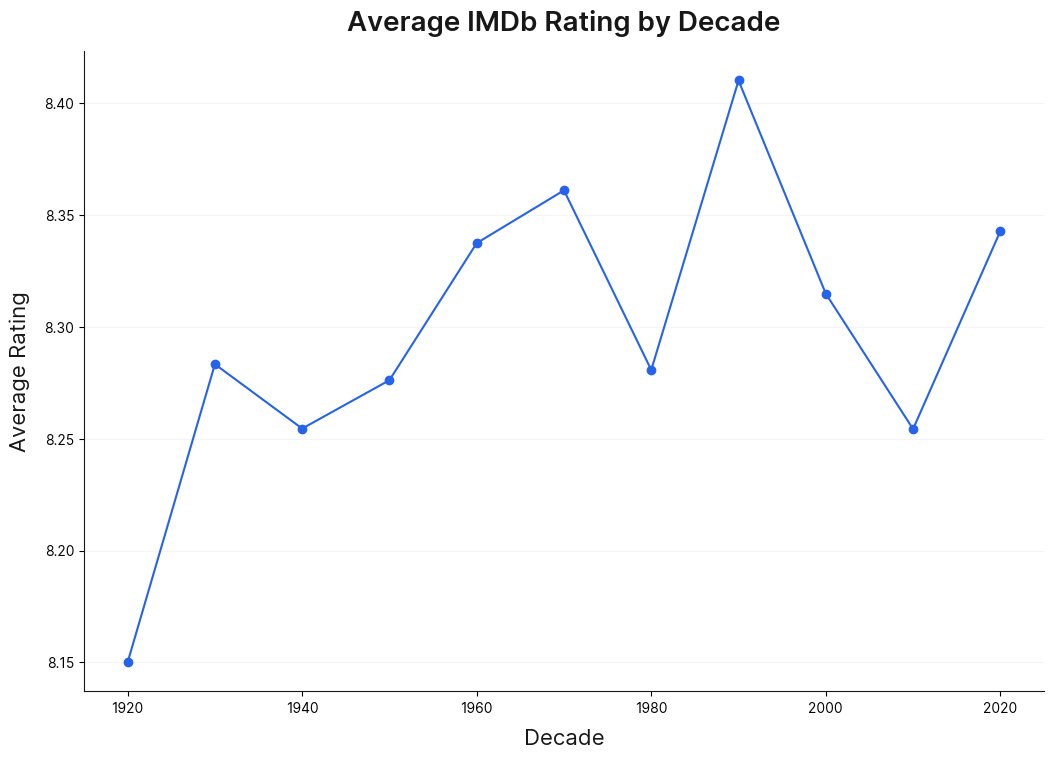
* **Key Findings:** The chart visually confirms that the highest-rated movies are consistently above a **9.0 rating**, with **"The Shawshank Redemption"** leading at **9.3**. The differences in rating among the top few movies are minimal, indicating a small cluster of critically acclaimed films.
* 

## **4. Analytical Questions and Insights**

This section addresses specific questions using further data analysis and visualizations.

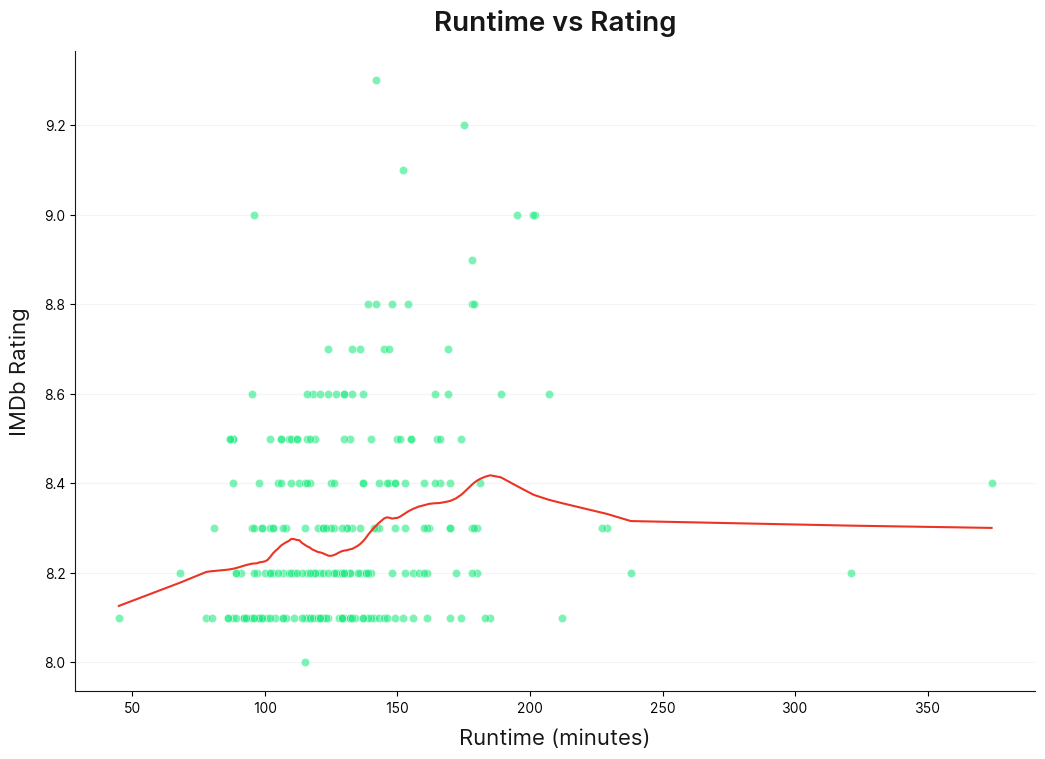
### **4.1. Which decade had the highest average IMDb ratings?**

| **Decade** | **Average Rating** |
| --- | --- |
| **1990** | **8.54** |
| 1970 | 8.52 |
| 1960 | 8.49 |
| 2000 | 8.42 |
| 1950 | 8.39 |

* **Answer:** The **1990s** (1990-1999) had the **highest average IMDb rating of 8.54**. This suggests that, on average, the movies released in this decade within the Top 250 list are rated slightly higher than those from other decades. The 1970s and 1960s follow very closely.
* 

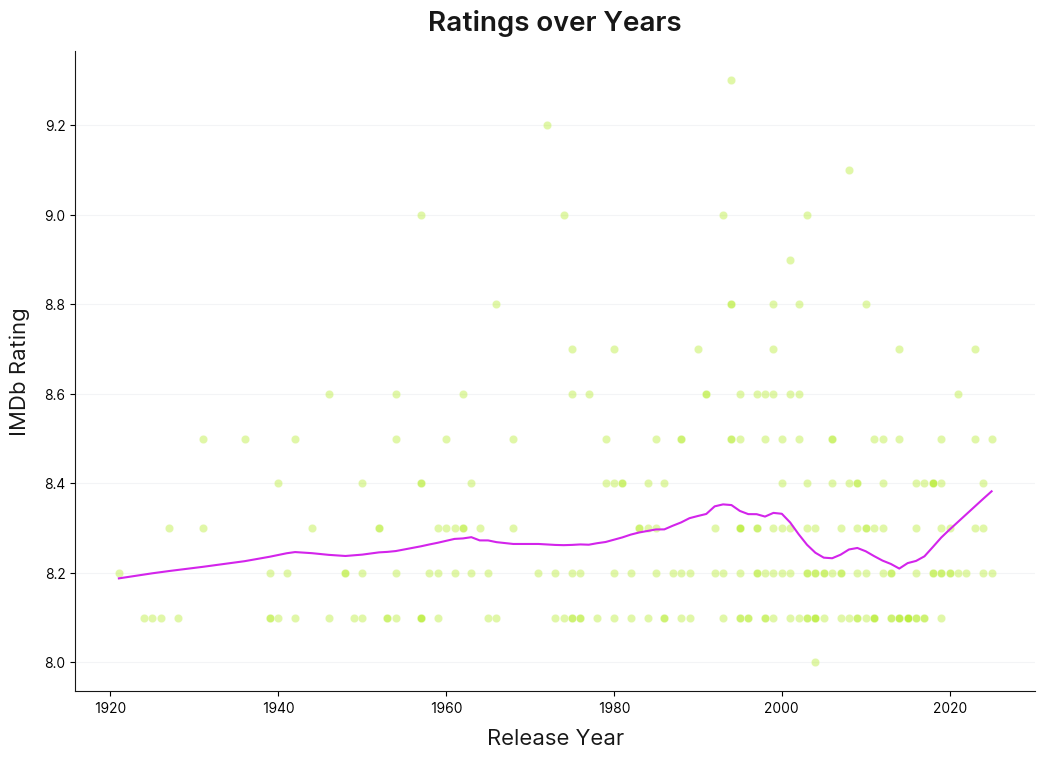
### **4.2. What’s the relationship between runtime and ratings?**

The scatter plot below explores the correlation between a movie's runtime and its IMDb rating.

* **Observation:** There appears to be a **weak positive correlation** between runtime and rating. While movies with the absolute highest ratings (9.0+) have runtimes generally between 140 and 200 minutes, the vast majority of highly-rated movies (8.1 to 8.5) are spread across a wide range of runtimes, from under 90 minutes to over 200 minutes.
* **Conclusion:** **Longer movies are not guaranteed higher ratings**, but the peak-rated movies tend to be longer than the median runtime.
* 

### **4.3. Are recent movies rated lower/higher than older ones?**

The line plot below shows the trend of the average IMDb rating over the years.

* **Observation:** The plot displays significant **volatility** in the average rating per year, as this is a limited sample of top movies (only 250). However, a general downward trend is visible:
  + Movies from the **1940s to the 1970s** show some of the highest average yearly ratings.
  + In recent decades (post-2000), the average yearly rating shows more fluctuation but generally stays in the lower part of the dataset's range (around 8.1 to 8.3), compared to some of the peak years in earlier decades.
* **Conclusion:** Based on the yearly average within this top-250 dataset, **older movies** (pre-1980) appear to be rated **slightly higher** on average than more recent ones, although this trend is not linear.
* 

### **4.4. Which year dominates the top list?**

* **Analysis:** By counting the number of movies per year present in the dataset:

| **Year** | **Count of Movies** |
| --- | --- |
| **1995** | **6** |
| 1994 | 5 |
| 1957 | 4 |
| 1971 | 4 |
| 2004 | 4 |

* **Answer:** The year **1995** has the **highest number of movies.**

## **5. Summary and Next Steps**

### **5.1. Summary of Insights**

* **Ratings Distribution:** Ratings are heavily concentrated in the **8.1 to 8.4** range, as expected from a filtered list of top-rated movies **(6)** in the IMDb Top 250 list. This suggests 1995 was a particularly strong year for critically successful films that earned enduring high ratings.ed movies.
* **Decadal Performance:** The **1990s** yielded the highest average-rated movies on the list (Avg. Rating: 8.54).
* **Runtime vs. Rating:** A **weak positive correlation** exists; while highly-rated movies tend to be longer, runtime is not a definitive predictor of a high rating.
* **Temporal Trend:** Older movies (pre-1980) have a slightly higher average rating than recent movies within this dataset.
* **Peak Year:** The year **1995** features the most movies (6) in this IMDb Top 250 compilation.