IMDb Top 250 Movies Data Analysis Documentation

1. Objective

The goal of this project is to extract, analyze, and derive insights from the IMDb Top 250 Movies dataset. The dataset is created by scraping a locally saved HTML file of the IMDb Top 250 Movies webpage. The data extracted includes movie title, release year, IMDb rating, and content rating. The final deliverables include data visualization, SQL-based querying, KPI metrics, and business insights through exploratory data analysis (EDA).

2. Data Collection

Source: IMDb Top 250 Movies HTML file (saved locally as "IMDb Top 250 Movies.html")  
  
Fields Extracted:  
- Title (string)  
- Year (int)  
- Rating (float)  
- Content Rating (string)  
  
Tools Used:  
- Python (BeautifulSoup, CSV)  
- Pandas (for data manipulation)  
- Matplotlib / Seaborn (for visualization)  
- SQLite / SQLAlchemy (for SQL queries)

3. Data Cleaning & Processing

- Removed leading index from movie titles (e.g., "1. The Shawshank Redemption" to "The Shawshank Redemption").  
- Converted year to integer.  
- Converted rating to float.  
- Handled missing content ratings with "Not Rated".  
- Saved final structured dataset in both CSV and JSON formats.

4. Exploratory Data Analysis (EDA)

Descriptive Statistics:  
- Total Movies: 250  
- Year Range: Oldest - 1921, Newest - 2023  
- Average Rating: ~8.6  
  
Top 5 Movies by Rating:  
- The Shawshank Redemption (9.3)  
- The Godfather (9.2)  
- The Dark Knight (9.0)  
- The Godfather Part II (9.0)  
- 12 Angry Men (9.0)  
  
Rating Distribution:  
- Most ratings range from 8.0 to 9.3  
- Very few movies below 8.0 or above 9.3  
  
Content Rating Distribution:  
- R: 90+ movies  
- PG-13: 80+ movies  
- PG: 40+ movies  
- Not Rated: ~10  
  
Yearly Trend:  
- More top-rated movies from the 1990s to 2010s  
- Classic hits from the 1950s to 1970s still retain high ranks

5. KPI Metrics

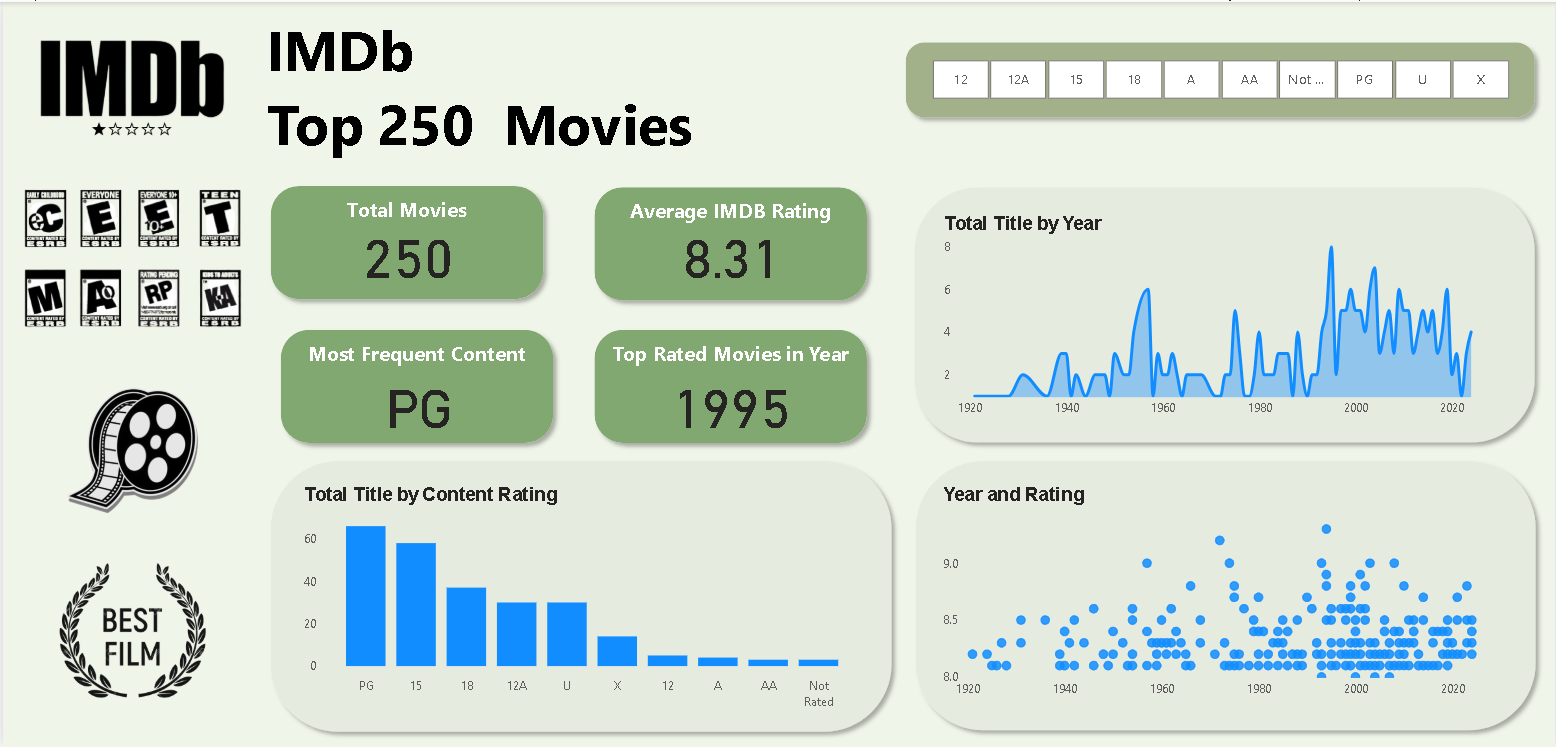
|  |  |
| --- | --- |
| KPI | Value |
| Total Movies | 250 |
| Average IMDb Rating | 8.31 |
| Most Frequent Content Rating | PG |
| Year with Most Top-Rated Movies | 1995 |

These KPIs provide a quantitative overview of dataset performance and viewer preferences.

6. SQL Queries for Business Insights

1. Top 10 Highest Rated Movies:  
SELECT Title, Rating FROM movies ORDER BY Rating DESC LIMIT 10;  
  
2. Count by Content Rating:  
SELECT Content\_Rating, COUNT(\*) as Count FROM movies GROUP BY Content\_Rating ORDER BY Count DESC;  
  
3. Average Rating by Decade:  
SELECT (Year / 10) \* 10 AS Decade, AVG(Rating) as Avg\_Rating  
FROM movies GROUP BY Decade ORDER BY Decade;  
  
4. Count of Movies Released Each Year:  
SELECT Year, COUNT(\*) as Movie\_Count FROM movies GROUP BY Year ORDER BY Year;

7. Data Visualization

Charts Created:  
- Bar Chart: Number of Movies by Content Rating  
- Line Chart: Number of Movies Released per Year  
- Scatter Plot: Year vs Rating

These visuals provide insights into trends and patterns in movie ratings, popularity over time, and audience targeting through content ratings.

8. Business Insights

- Movies rated "R" dominate the IMDb Top 250 list, indicating a strong adult viewer preference.  
- The 1990s and 2000s contributed significantly to the top-rated movies, suggesting a peak period for acclaimed cinema.  
- Average rating has remained consistently high across decades, proving timeless appeal.  
- Directors and producers targeting PG-13 and R audiences appear to have better chances of creating IMDb-ranked classics.

9. Conclusion

The IMDb Top 250 dataset reveals clear trends in cinematic excellence and audience preferences. Through structured data extraction, SQL analysis, KPI evaluation, and visual storytelling, we derived actionable insights on content rating strategies, temporal trends, and rating behavior. This analysis can support decisions in film marketing, content planning, and historical studies in cinema.  
  
Tools Used: Python, Beautiful Soup, Pandas, MySQL , Matplotlib, Seaborn  
  
Author: Bajithu R.