COURSE WORK: PROGRAMMING FOR DATA SCIENCE 2024.2 BATCH

MSc Data Science: Coventry University UK

Author: B M J N Balasuriya
Index: COMScDS242P-009

Exploratory Data Analysis on Top IMDb Top 250 Movies.

Introduction

The IMDb Top 250 Movies dataset is a collection of the highest-rated films as determined by IMDb users.

This dataset includes a variety of attributes for each movie, such as:

- Title: The name of the movie.
- Year: The release year of the movie.
- Rating: The IMDb rating, which is an average of user ratings.
- Genre: The genre(s) of the movie.
- Director(s): The director(s) of the movie.
- Box Office Revenue: The gross worldwide revenue of the movie.
- Lead Actors: The main actors in the movie.

Github repository: https://github.com/j2damax/eda_top_us_song

Follow these steps given the *README.md* file to setup the project.

1.Data Loading and Preprocessing:

data_scraper.py:

The data_scraper.py script automates the process of extracting detailed information about the IMDb Top 250 Movies. Key features include:

Selenium WebDriver:

- Navigates to the IMDb Top 250 Movies page.
- Loads dynamic content.
- The script visits each movie's individual page to extract additional details such as genre, directors, box office revenue, and lead actors.

BeautifulSoup:

- Parses HTML content.
- Extracts movie details.

Data Extraction:

- For each movie, the script extracts the title, release year, rating, and link to the movie's individual page

Data Storage:

- The extracted data is stored in a Pandas DataFrame and saved as a CSV file (imdb_top_movies.csv)

2. Statistical Analysis:

- This data set contains 250 data rows and 8 data columns about movies.
- When data scraping found that 4 movies doesn't included revenue and after doing a google search, found that these were some missing data on the data set except one movie.
- "Jai Bhim" movie did not open in theatres so there are no box office numbers to support its popularity.
- "Hamilton", "Klaus", "Drishyam" has actual revenue and replaced these data manually into the dataset.
- The **ratings** of the movies are generally high and consistent, with most movies having ratings around **8.31**(mean). The small standard deviation indicates that there is little variation in the ratings. (mean = 8.31, median = 8.20, std = 0.23)
- The revenue data shows a significant skew, with a few movies earning exceptionally high revenues, which pulls the mean up. The large standard deviation reflects the wide range of revenue values. (mean = 229707813.30, median = 62514489.00, std = 371066078.43)
- The votes data also shows a right-skewed distribution, with some movies receiving a very high number of votes. The large standard deviation indicates a wide variation in audience engagement. (mean = 736296.75, median = 627500.00, std = 587452.30)
- The boxplot reveals several outliers, representing movies with exceptionally high revenues compared to the majority.
- The revenue data is highly skewed, with most movies earning significantly less than the few blockbusters that dominate the box office.
- The interquartile range (IQR) indicates a wide spread of revenue values, reflecting the variability in movie earnings. (IQR 314143988.0)
- The outliers correspond to blockbuster movies that generate disproportionately high revenues.
- Most movies fall within a lower revenue range, as indicated by the concentration of data points near the lower end of the boxplot.
- The ratings are concentrated within a narrow range, typically between 8.0 and 9.5, indicating that the dataset focuses on highly-rated movies.
- There are a few outliers with exceptionally high or low ratings, but they do not significantly affect the overall distribution. (Q1 (25th percentile): 8.1, Q2 (Median): 8.2, Q3 (75th percentile): 8.4)
- The small interquartile range (IQR) suggests that the ratings are consistent, with most movies receiving similar high ratings. (IQR=0.3000000000000007)

3. Data Visualization:

Distribution of Revenue

- The distribution of revenue appears to be highly skewed, with a majority of movies earning significantly lower revenue compared to a few high-grossing movies.
- There are noticeable outliers in the data, representing blockbuster movies with exceptionally high revenue.
- The majority of movies fall within a lower revenue range, indicating that high revenue is not common across all movies.
- The presence of a long tail in the distribution suggests that revenue varies widely among movies, with a small number of movies contributing disproportionately to the total revenue.

Distribution of Ratings

- The ratings appear to follow a relatively normal distribution, with most movies clustered around a central value.
- The majority of movies have high ratings, indicating that the dataset primarily consists of well-rated movies.
- The range of ratings is narrow, typically between 8.0 and 9.5, suggesting that the dataset focuses on top-rated movies.
- There are few outliers with exceptionally high or low ratings, indicating consistency in the quality of movies in the dataset.

Distribution of Movie Release Years

- The distribution of movie release years show a skewed pattern, with more movies being released in recent decades compared to earlier years.
- The distribution might highlight historical trends in the film industry, such as the rise of blockbuster movies or the impact of technological advancements.
- Older movies (e.g., from the 1950s or earlier) might appear as outliers, representing classics or historically significant films.

Trend of Revenue Over Time

- There is a noticeable upward trend in movie revenues over time, indicating that movies released in recent years tend to generate higher revenues compared to older movies.

- The trend highlights the emergence of blockbuster movies in recent decades, with some movies achieving exceptionally high revenues.

- he increase in revenue over time may reflect advancements in technology, marketing, and global distribution, which have expanded the audience reach and revenue potential of movies.

- The rise in revenue could also be linked to the growing popularity of certain genres, such as superhero movies and franchises, which dominate the box office in recent years.

4. Movie and Director Trends:

- Top 5 recurrent directors are “Robert De Niro”(9), “Steven Spielberg”(8), “Christopher Nolan”(8), “Stanley Kubrick”(7) and “Martin Scorsese”(7) over the years.

- Top 5 recurrent actors are “Robert De Niro”(9), “Harrison Ford”(6), “Leonardo DiCaprio”(6), “Tom Hanks”(6) and “Clint Eastwood”(5) over the years.

- Directors like Robert De Niro, Steven Spielberg, and Christopher Nolan frequently appear in the dataset with high-rated and high-revenue movies. Their consistent presence in top-grossing and critically acclaimed movies highlights their influence on movie success.

- Actors like Robert De Niro, Leonardo DiCaprio, and Tom Hanks frequently appear in the dataset. Their involvement in movies often correlates with high ratings and significant box office success.

- There is a noticeable peak in the release of superhero movies in recent years, indicating a surge in production and demand for this genre. ‘(Action', 'Superhero', ‘Thriller’). This likely due to the release of major franchises or blockbuster hits.

5. How different movie genres have evolved over the years:

- Certain genres, such as Drama, Adventure, and Epic, consistently dominate over time, reflecting their enduring popularity and appeal.

- Some genres, like Superhero and Sci-Fi, gain prominence in recent decades, driven by advancements in technology and the rise of blockbuster franchises.

- Genres like Drama and Thriller maintain a consistent presence, highlighting their universal appeal across different eras.

- Sci-Fi movies had a minimal presence in the early decades, with only one or two movies per decade in the 1920s and 1960s.

- The 1980s and 2010s saw the highest number of Sci-Fi movies, indicating a peak in the genre's popularity during these decades.

6. Correlations and Discoveries:

Revenue vs. Votes

- There is a noticeable positive correlation between the number of votes and revenue. Movies with higher audience engagement (votes) tend to generate higher revenue.
-
- Many movies with lower votes are clustered around lower revenue, indicating limited audience reach or niche appeal.

Rating vs. Votes

- There is a weak positive correlation between ratings and votes. Movies with higher ratings tend to receive more votes, indicating greater audience engagement.

- Most movies are clustered around high ratings (8.0–9.0), reflecting the dataset's focus on top-rated movies.

- Movies with average ratings (around 8.0) still receive a substantial number of votes, indicating that audience preferences are not solely driven by ratings.

8. Conclusion:

Interesting Observations

** **Consistency in Ratings**:
Ratings for top movies remain consistently high, indicating a focus on critically acclaimed films in the dataset.

** **Blockbuster Effect**:
A small number of movies contribute disproportionately to total revenue, reflecting the dominance of blockbuster franchises.

** **Cultural Impact**:
The rise of Superhero and Sci-Fi genres highlights their cultural significance and audience demand.

** **Collaborations**:
Repeated collaborations between directors and actors, such as Steven Spielberg and Tom Hanks, often result in successful movies.

The analysis highlights the significant influence of directors, actors, and genres on movie success. While revenue is driven by factors like marketing and franchise appeal, ratings reflect consistent quality among top movies. The dataset provides valuable insights into the evolution of cinema, audience preferences, and the factors contributing to critical and commercial success.