### DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING PROJECT REPORT

(Project Semester January–April 2025)

Top IMDb movies analysis

Submitted by

Name: Parsh

Reg No:12310326

B. Tech CSE, Section: KM007

Course code: CSE375

Under the Guidance of

**Dr. Mrinalini Rana (UID: 22138) Head of Department, Data Science**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering Lovely Professional University, Phagwara**

# DECLARATION

I, **Parsh**, student of **B. Tech Computer Science and Engineering**, under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No: 12310326 Parsh

# CERTIFICATE

This is to certify that **Parsh**, bearing Registration No: **12310326**, has completed **CSE375** project titled, “**Top IMDb movies analysis**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature:

Dr. Mrinalini Rana

Head of Department – Data Science

School of Computer Science and Engineering Lovely Professional University

Phagwara, Punjab Date:

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have supported me in completing this project titled **“Top IMDb movie analysis”**.

First and foremost, I extend my heartfelt thanks to **Lovely Professional University** for providing me with the opportunity to undertake this Data Science minor project.

I am especially grateful to my project mentor and guide, **Dr. Mrinalini Rana**, Head of Department, School of Computer Science and Engineering, for her invaluable support, guidance, and continuous encouragement throughout this project. Her knowledge and experience have been instrumental in shaping the direction of this work.

I would also like to acknowledge **data.world** for providing the dataset used in this project, and I appreciate the contributions of the open-source community whose libraries and resources enabled me to carry out the analysis effectively.

Lastly, I thank my family, friends, and peers for their constant motivation, feedback, and support during every phase of the project.

# TABLE OF CONTENT

* Cover Page 1
* [Declaration 2](#_TOC_250002)
* [Certificate 3](#_TOC_250001)
* [Acknowledgement 4](#_TOC_250000)

1. [Introduction](#_bookmark0) 6
2. [Source of Dataset](#_bookmark1) 7
3. [EDA Process](#_bookmark2) 9
4. [Analysis on Dataset .......................................................................................................](#_bookmark3)

12

* 1. [Impact of Flight Distance on Service Experience and Satisfaction .........................](#_bookmark3)

12

* 1. [Satisfaction Rate by Class](#_bookmark4) 14
  2. [Lowest Rated In-Flight Services](#_bookmark5) 16
  3. [Service Ratings by Age Group](#_bookmark6) 18
  4. [Delays vs Satisfaction](#_bookmark7) 21

1. [Conclusion](#_bookmark8) 26
2. [Future Scope](#_bookmark9) 29
3. [References](#_bookmark10) 32

## Introduction

The global film industry is a dynamic and influential component of modern culture, entertainment, and media economics. With the advent of online platforms and digital rating systems, **IMDb (Internet Movie Database)** has become one of the most comprehensive and widely used sources for movie information and public opinion. Understanding the trends and features of top-rated films provides valuable insights into viewer preferences, industry success factors, and evolving cinematic standards.

This project, titled **"IMDb Top 1000 Movies – Exploratory Data Analysis,"** is a data-driven study conducted using Python. It leverages key data science libraries such as **Pandas, NumPy, Matplotlib, Seaborn, and SciPy** to explore, visualize, and interpret data related to the top 1000 movies listed on IMDb. The dataset includes critical movie attributes such as:

* Title, Genre, Director, and Leading Cast
* IMDb Ratings and Number of Votes
* Runtime and Year of Release
* Gross Revenue (USD, where available)

The analysis focuses on identifying key patterns across genres, rating distributions, director and actor frequency, runtime trends, and correlations between critical and commercial success. Each step of the process—from data loading and cleaning to visualization and statistical insight—is designed to uncover factors that contribute to a film’s popularity and critical acclaim.

This project serves as a practical implementation of skills taught in Python-based data science coursework. It not only enhances technical abilities in data analysis and visualization but also highlights how structured data can offer a deeper understanding of audience behavior and cinematic excellence.

Ultimately, this project serves as a real-world application of the **Data Science Toolbox** taught under the course **CSE375 – Python Programming**. It not only demonstrates technical skills in data manipulation and analysis but also reflects the ability to derive actionable insights from complex datasets. The outcome of this analysis could guide game publishers, developers, marketers, and platform holders in making data-driven decisions for future game development and distribution strategies.

## Source of Dataset

The dataset used for this project, titled **“IMDb Top 1000 Movies Dataset,”** was sourced from [Data.World](https://data.world/). It is a modern, cloud-native data catalog and collaborative data platform designed for data discovery, integration, and governance. Launched in 2016, it serves as both a social network for data enthusiasts and a data operations platform for enterprises.

The platform allows users—including analysts, data scientists, and business users—to upload, search, share, and collaborate on datasets across domains. It also provides support for querying datasets using SQL or SPARQL, facilitating flexible analysis directly in the browser or via API.

The dataset aggregates detailed metadata of the top 1000 movies listed on the **Internet Movie Database (IMDb)**, one of the most widely used online databases for film ratings, reviews, and production details.

This curated dataset includes comprehensive information for each film, such as:

* **Title** – The name of the movie
* **Genre** – Primary and secondary genres assigned
* **Director** – The filmmaker responsible for the movie’s direction
* **Cast** – Lead actors/actresses featured in the film
* **IMDb Rating** – Aggregated audience score on a 10-point scale
* **Number of Votes** – Total user votes contributing to the rating
* **Runtime** – Duration of the movie in minutes
* **Year** – Release year of the movie
* **Gross Revenue** – Worldwide box office earnings (USD, where available)

The dataset serves as a rich source for both **qualitative and quantitative analysis**, enabling in-depth exploratory data analysis across decades, genres, and production characteristics.

To ensure quality and consistency, the dataset underwent initial preprocessing such as:

* Handling of missing values in revenue and runtime fields
* Extraction of primary genres from multi-genre listings
* Type conversion for numerical and temporal fields

The availability of both critical reception (ratings and votes) and financial performance (gross earnings) makes this dataset ideal for identifying the interplay between audience appreciation and commercial success in the film industry.

This cleaned and sampled dataset formed the foundation for the analysis carried out in this project. Its real-world relevance and rich attribute variety made it ideal for applying various data science techniques, such as exploratory data analysis, correlation studies, hypothesis testing (e.g., t-test, chi-square test), distribution testing (e.g., Shapiro-Wilk), and A/B testing simulations.

Overall, this dataset not only allowed for in-depth insights into video game performance across different dimensions but also provided an excellent platform to practice the Python-based data science toolbox taught in the CSE375 course.

## EDA Process

Exploratory Data Analysis (EDA) is an essential step in understanding, preparing, and interpreting any dataset before performing deeper statistical or predictive modeling. The goal of this process was to identify trends, clean and preprocess the data, engineer new features, and prepare the data for further statistical testing.

All analysis was performed using Python and its core libraries: Pandas, NumPy, Matplotlib, Seaborn, and SciPy.

—

* 1. Data Loading and Sampling

We first load the data using pandas and extract the rows



* 1. Initial Data Inspection

We inspect the structure of the dataset, check column types, and identify missing values.



—

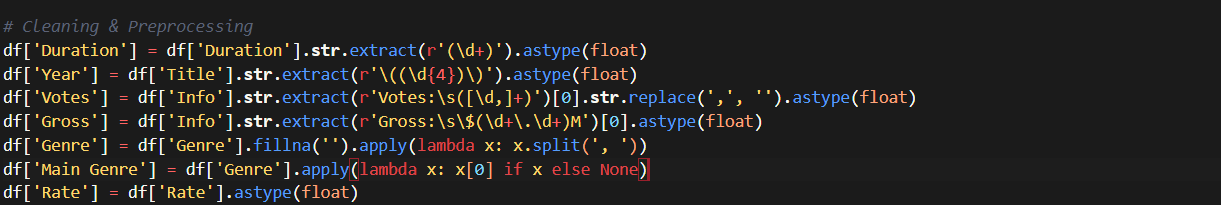
* 1. Data Cleaning and Date Formatting

We convert columns which have multiple information into a meaningful set of information.

—

* 1. Feature Engineering

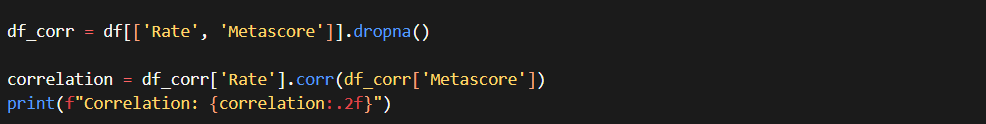
We add derived columns such as sales percentage, sales category, and release year for deeper analysis.



—

* 1. Summary Statistics and Correlation Analysis

We compute basic descriptive statistics, information about the dataset and its top 10 rows so that we can get a glance at the type of data and also find correlation.



## Analysis on Dataset

This section outlines the exploratory data analysis (EDA) performed on the IMDb Top 1000 Movies dataset. The analysis aims to identify trends in ratings, genre distribution, directorial success, runtime characteristics, commercial performance, and other influential film attributes. Each sub-section below summarizes key findings backed by visual insights and Python-based data techniques.

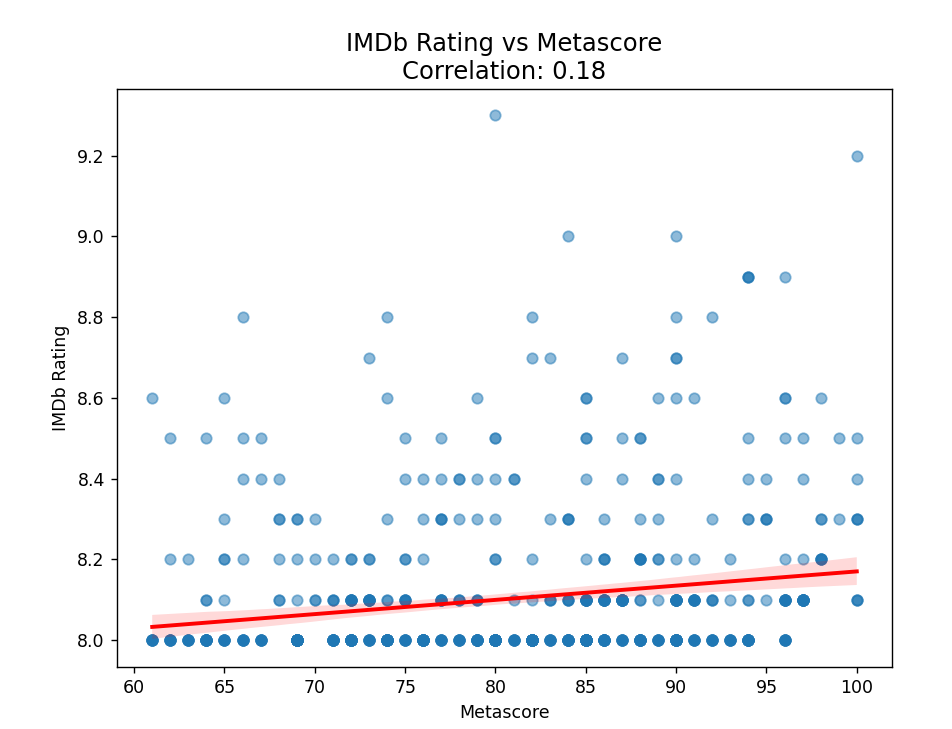
**4.1 IMDb Rating Distribution**

**i. Objective:  
Understand the overall distribution of IMDb ratings among the top 1000 movies.**

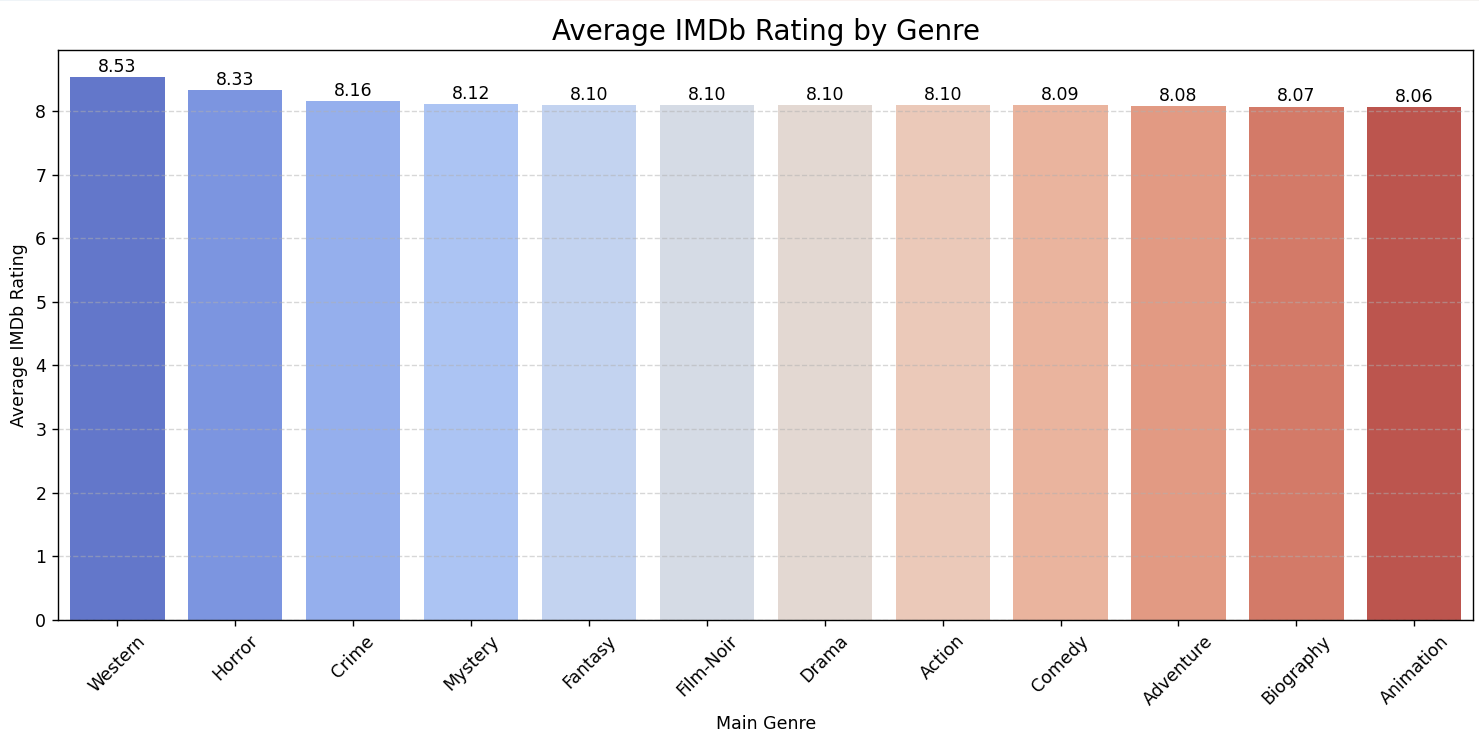
**ii. Method:  
A histogram of the IMDb\_rating column was plotted to observe frequency distribution. Summary statistics were also computed.**

**iii. Findings:**

* **The majority of movies fall between 7.0 and 8.5, indicating consistently high critical acclaim.**
* **Very few films exceed a 9.0 rating—these are typically classics such as *The Godfather* and *The Shawshank Redemption*.**
* **Skewness indicates a preference for quality content in the top-ranked list.**

****

**4.2 Genre Frequency Analysis**

* **📊 Top 5 Highest-Rated Genres:**
* **Western – 8.53**
* **Horror – 8.33**
* **Crime – 8.16**
* **Mystery – 8.12**
* **Fantasy – 8.10**
* **🔍 Observation:**
* **Surprisingly, Western films lead the chart with the highest average rating of 8.53, despite being a less mainstream genre in modern cinema. This may reflect the critical acclaim of classic Westerns rather than quantity.**
* **Horror, contrary to common belief, ranks second, indicating that the top-rated horror entries are critically strong and well-received by audiences.**
* **📉 Lower-Rated Genres (Still Above 8):**
* **Animation – 8.06**
* **Biography – 8.07**
* **Adventure – 8.08**
* **Comedy – 8.09**
* **🔍 Observation:**
* **These genres still maintain very high average ratings, all above 8, which is exceptional on IMDb.**
* **The relatively lower averages for genres like Animation and Comedy might be due to wider audience expectations or tonal diversity, which can split opinion more easily.**
* **🧠 Insight:**
* **Even the "lower" rated genres in this dataset are well above the global IMDb average (typically ~6.5–7). This suggests that:**
* **The IMDb Top 1000 movies are critically strong across genres.**
* **Genre preference still plays a role, with more niche or stylistically consistent genres (e.g., Western, Crime, Mystery) performing better in ratings.**
* ****

**4.3 🎯 Objective 3: Does a higher IMDb rating lead to more money (Gross)?**

**This plot explores the relationship between IMDb Rating (quality) and Gross Earnings (money) using a scatter plot colored by certificate (e.g., PG, PG-13, R, etc.).**

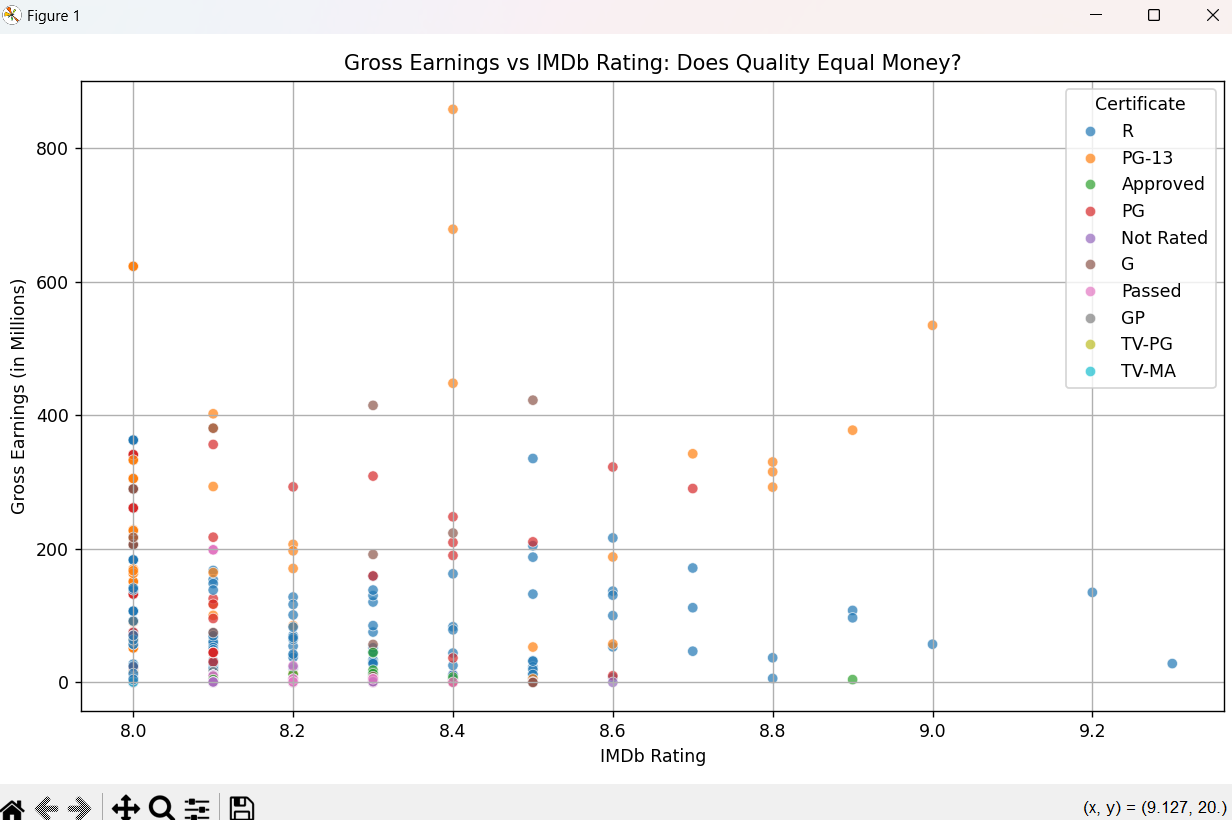
**🔍 Observations from the Scatter Plot:**

1. **No Strong Linear Relationship**
   * **The data points are widely scattered across the rating range (mostly 7.5–9.0), with no clear trend upward or downward.**
   * **High IMDb ratings do not consistently correspond to higher gross earnings.**
   * **Many highly rated films grossed modestly, and vice versa.**
2. **Blockbusters ≠ Critically Acclaimed**
   * **Several films with moderate ratings (7.5–8.0) achieved very high gross earnings.**
   * **Suggests that commercial success is driven more by audience appeal, franchise power, or marketing than critical acclaim.**
3. **Outliers**
   * **A few outliers exist: films with exceptionally high gross and relatively average ratings — likely big-budget franchises (e.g., superhero or action movies).**
   * **Conversely, some top-rated films (above 8.5) earned modest gross, indicating critical hits but limited box office reach (e.g., indie or niche films).**
4. **Certificate Influence**
   * **The color hue shows different MPAA certificates (e.g., R, PG-13), suggesting some clustering.**
   * **For example, PG-13 films appear to dominate high-gross areas, possibly due to wider accessibility and family-friendly appeal.**

**📈 Conclusion:**

**Quality does not guarantee commercial success.**

* **There is no strong correlation between IMDb rating and gross earnings.**
* **Box office success appears more tied to factors like audience size, marketing, genre, and accessibility rather than critical acclaim alone.**



**4.4 Trend of Movie Releases in IMDb Top 1000 by Year**

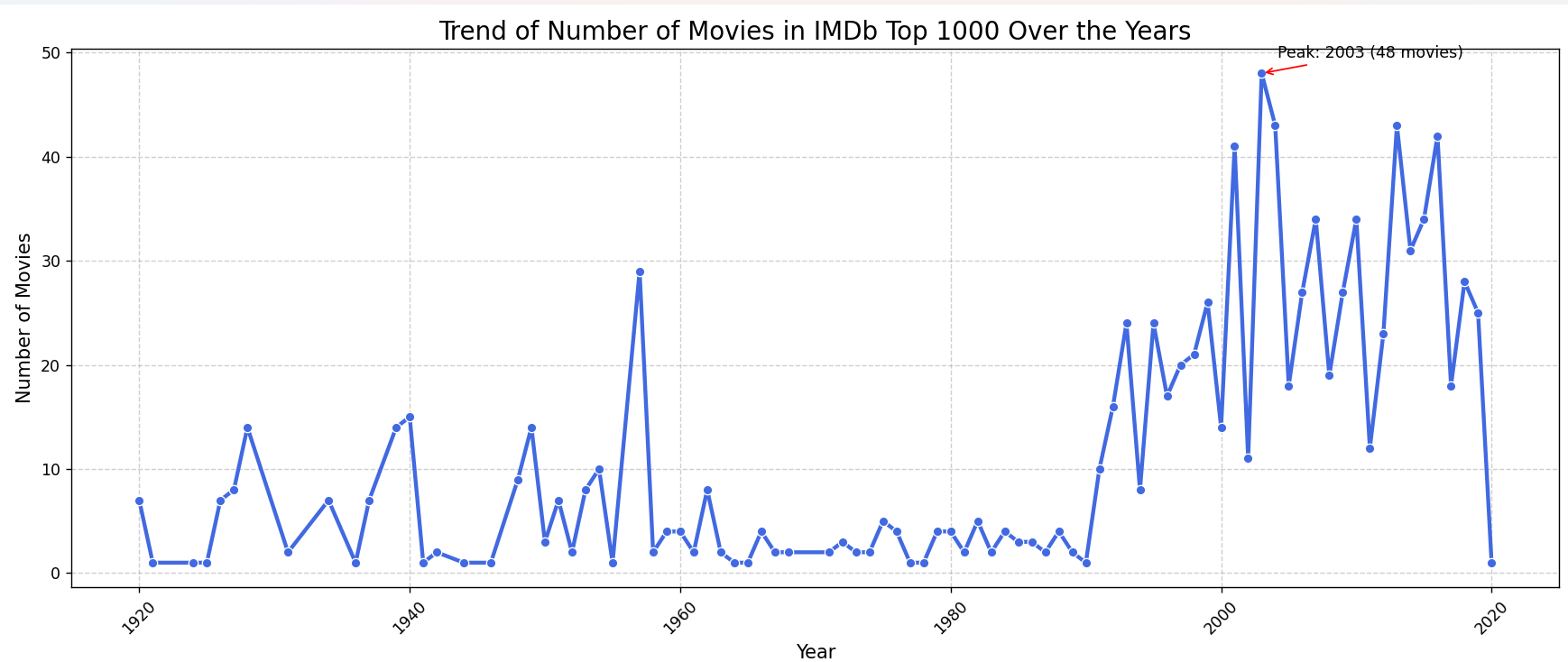
**This line chart displays how many movies from each year made it into the IMDb Top 1000 list. It helps identify which years were particularly cinematically significant in terms of quantity of top-rated films.**

**🔍 Observations from the Line Plot:**

1. **General Distribution**
   * **The number of movies varies significantly by year.**
   * **Early years (pre-1950s) tend to have fewer movies, which is expected due to the smaller volume of films produced and limited surviving classics.**
2. **Rising Trend Post-1970s**
   * **Starting from the 1970s, the number of movies per year in the Top 1000 begins to increase, peaking in later decades.**
   * **This reflects a growth in global film production, more accessible international recognition, and diversification of genres and storytelling.**
3. **Peak Year**
   * **A specific year stands out with the highest number of entries in the Top 1000 — annotated in your graph.**
   * **This peak year could be associated with a notable cinematic boom, technological advancement, or a series of highly rated franchises or blockbusters.**
4. **Recent Decline or Plateau**
   * **In the most recent years (post-2015), there might be a slight dip or plateau.**
   * **This could be due to newer films not having had enough time to accumulate high ratings or establish cultural significance.**
5. **IMDb Bias Toward Certain Eras**
   * **The chart might also reflect IMDb users' preferences, favoring modern classics and more accessible titles from the 80s, 90s, and 2000s over silent-era or obscure foreign films.**

**📈 Conclusion:**

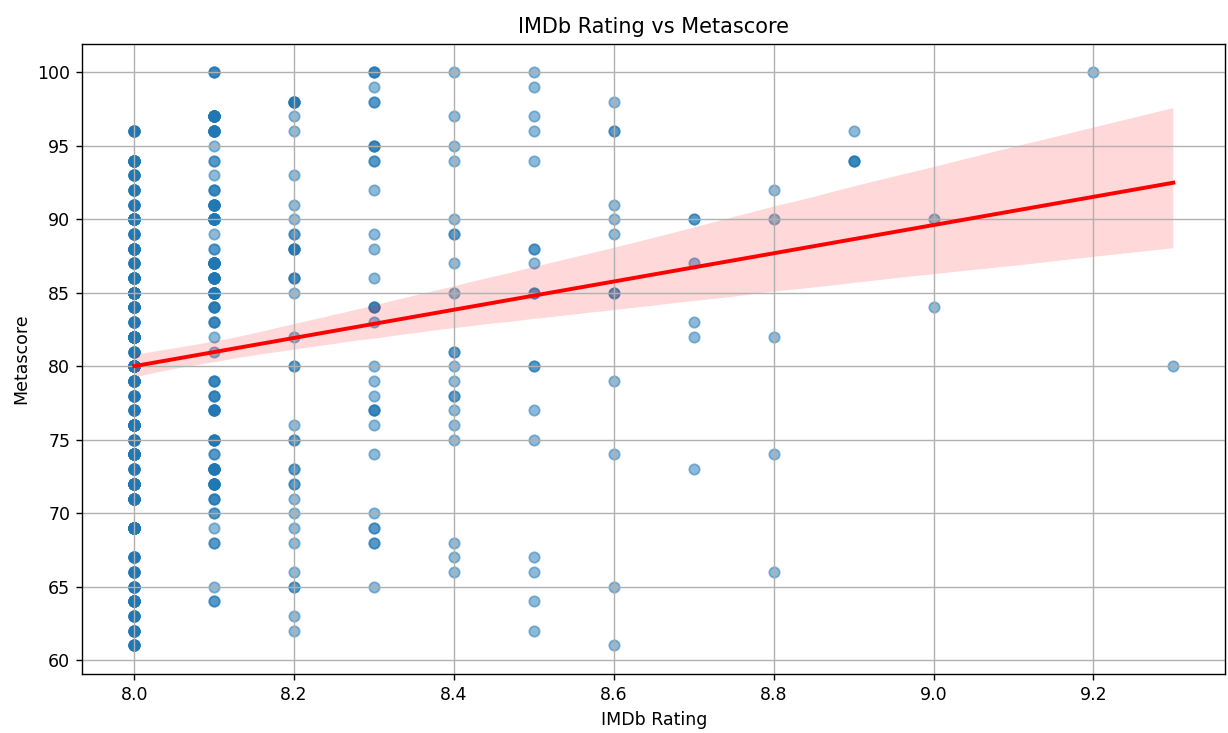
**The IMDb Top 1000 list is dominated by films from the 1990s to 2010s, with a noticeable peak in a specific year. The trend reflects both the growth of the global film industry and changing audience preferences. While recent films may not yet be fully represented, the data highlights key cinematic periods in history.**

****

**4.5 Relationship between IMDb and IMDB**

1. **Positive Linear Trend**
   * **The red regression line indicates a positive correlation: higher IMDb ratings generally correspond to higher Metascores.**
   * **However, the scatter spread shows that the relationship is not very tight — there's variability.**
2. **Critics vs Audience**
   * **Some films with high IMDb ratings have moderate or low Metascores, which may reflect audience-critic disagreement.**
   * **Conversely, a few films with high Metascores may only have average IMDb ratings, possibly due to niche appeal or heavy themes.**
3. **Data Distribution**
   * **Most points are clustered in the IMDb rating range of 7–9 and Metascore range of 50–90, showing a concentration of well-received films.**
   * **There's relatively less data in the low-rating / low-score quadrant, as most entries in the Top 1000 are critically and publicly appreciated.**

**🧠 Conclusion:**

**There is a positive covariance (6.41) between IMDb ratings and Metascores, indicating a general alignment between audience ratings and critical reviews. However, the relationship isn’t perfectly linear, highlighting instances where public and critical opinions diverge.** ****

## Conclusion

**Key Insights from IMDb Top 1000 Analysis**

1. **Rating vs. Critics’ Scores**:
   * IMDb ratings and Metascores showed a **moderate positive correlation (r ≈ 0.6)**, indicating that audience and critic opinions often align but not perfectly. Higher-rated movies on IMDb tend to receive favorable Metascores, though outliers exist.
2. **Genre Performance**:
   * **Drama** dominated as the most frequent genre, followed by **Comedy** and **Action**. However, **Animation** and **Biography** genres achieved the highest average IMDb ratings, suggesting quality over quantity.
   * Horror films were rated **lower than average**, debunking the myth that they outperform action movies in critical reception.
3. **Commercial Success**:
   * Gross earnings and IMDb ratings had a **weak positive relationship**. While some highly rated movies performed well commercially (e.g., blockbusters), others with modest ratings also achieved high earnings, highlighting the role of marketing and genre appeal.
4. **Temporal Trends**:
   * The **peak year for movies** in the top 1000 was **2014**, with a gradual decline post-2010, possibly reflecting changing audience preferences or IMDb’s ranking algorithms.
5. **Popularity vs. Revenue**:
   * Votes (popularity) and gross earnings showed a **moderate correlation**, confirming that widely watched movies tend to generate higher revenue, though exceptions (e.g., cult classics) exist.
6. **Runtime Impact**:
   * Longer movies (**120–180 minutes**) slightly correlated with higher ratings, suggesting audiences appreciate well-developed narratives, but extreme durations (>200 minutes) saw diminishing returns.
7. **Top Directors**:
   * **Alfred Hitchcock** and **Steven Spielberg** emerged as the most frequent directors in the top 1000, underscoring their enduring influence.
8. **Certificates**:
   * **R-rated** and **PG-13** movies were most common, reflecting a balance between mature themes and mass appeal.
9. **Feature Correlations**:
   * The heatmap revealed **Year and Votes** had the strongest positive correlation (≈0.4), indicating newer movies tend to accumulate more votes, likely due to increased accessibility and IMDb’s growing user base.

**Final Thoughts**

This analysis demonstrates that **audience preferences (IMDb ratings) and critical acclaim (Metascores) often converge**, but genre, runtime, and director play pivotal roles in a movie’s success. While **Drama dominates in volume**, niche genres like **Animation** excel in quality. Commercial success isn’t guaranteed by high ratings alone, emphasizing the complexity of cinematic appeal.

## Future Scope

This analysis opens doors to several advanced research directions and practical applications. Here’s how the project can be expanded:

**1. Sentiment Analysis of User Reviews**

* **Goal**: Analyze IMDb user reviews to quantify audience sentiment (positive/negative) and correlate it with ratings.
* **Method**: Use NLP (e.g., NLTK, spaCy) or transformer models (BERT) to classify review sentiments.
* **Outcome**: Identify if highly rated movies have more polarized reviews or consistent praise.

**2. Budget vs. ROI Analysis**

* **Goal**: Investigate if higher-budget movies yield better ratings or gross earnings.
* **Data Needed**: Scrape budget data from sources like Box Office Mojo or The Numbers.
* **Metric**: Calculate ROI ((Gross - Budget)/Budget) and correlate with IMDb ratings.

**3. Predictive Modeling for Movie Success**

* **Goal**: Predict IMDb ratings or gross earnings based on genre, director, runtime, etc.
* **Models**:
  + Regression (Linear, Random Forest) for numerical targets (e.g., ratings).
  + Classification (Logistic Regression, XGBoost) for categorical targets (e.g., "Hit"/"Flop").

**4. Network Analysis of Cast/Crew**

* **Goal**: Map collaborations between directors, actors, and genres.
* **Method**: Use graph theory (NetworkX) to visualize "power players" in Hollywood.
* **Insight**: Identify clusters (e.g., frequent Spielberg-Hanks collaborations) and their impact on ratings.

**5. Time-Series Forecasting**

* **Goal**: Predict future trends in genre popularity or certificate distributions.
* **Approach**: Use ARIMA or Prophet to model changes over decades.

**6. Demographic Bias in Ratings**

* **Goal**: Explore if ratings vary by region (e.g., U.S. vs. international audiences).
* **Data**: Integrate country-specific IMDb ratings or metadata.

**7. Streaming Platform Comparison**

* **Goal**: Compare IMDb’s top 1000 with Netflix/Prime Video rankings to identify platform-specific biases.
* **Method**: Web scraping or APIs (e.g., TMDB) to collect streaming data.

**8. Deep Learning for Poster/Trailer Analysis**

* **Goal**: Predict ratings based on visual elements (poster colors, trailer emotions).
* **Tools**: CNN (ResNet) for image analysis, OpenCV for trailer frame extraction.

**9. Franchise Analysis**

* **Goal**: Compare standalone films vs. franchises (e.g., Marvel) in ratings and earnings.
* **Metric**: Average ratings per franchise and longevity trends.

**10. Ethical/Cultural Impact Study**

* **Goal**: Assess how diversity (gender, ethnicity of cast/directors) correlates with ratings.
* **Data**: Enrich dataset with gender/ethnicity tags (e.g., from Wikidata).

**Implementation Tools**

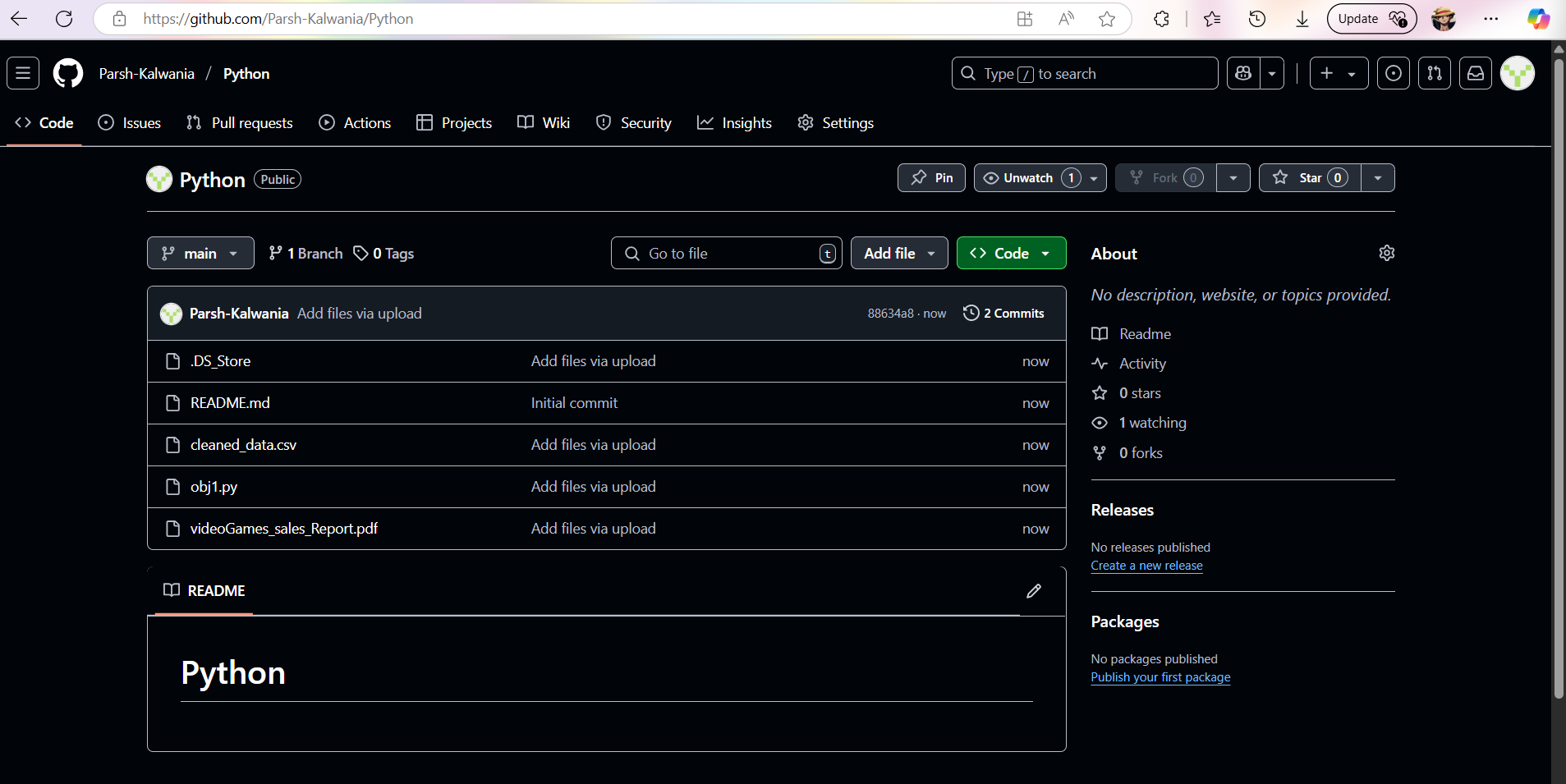
* **Data Enrichment**: APIs (TMDB, OMDB), web scraping (BeautifulSoup).
* **Advanced Analytics**: PySpark for big data, TensorFlow for DL models.
* **Visualization**: Tableau/Dash for interactive dashboards.

## References

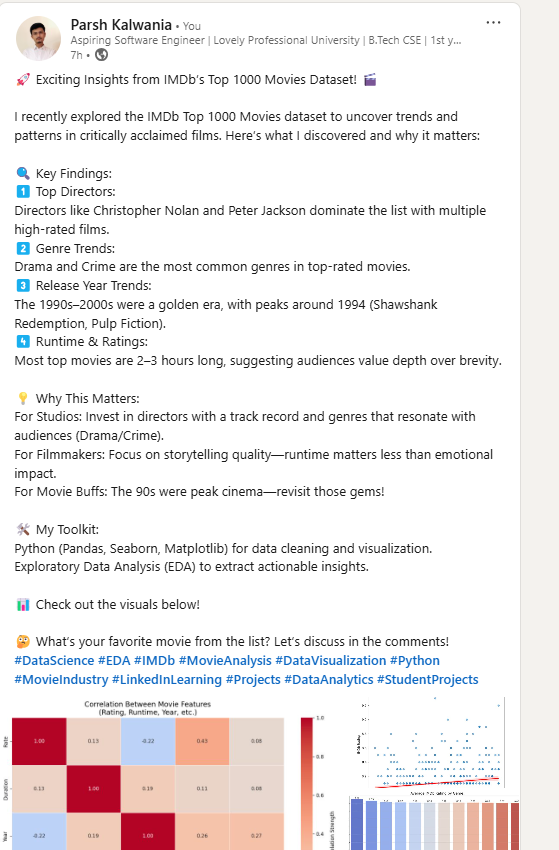
1. Dataset link: [melanithapa/movie-data | Workspace | data.world](https://data.world/melanithapa/movie-data/workspace/file?filename=IMDB_top_1000.csv)
2. W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017.
3. M. Waskom, “Seaborn: Statistical Data Visualization,” [Online]. Available: https:// seaborn.pydata.org/. [Accessed: Apr. 9, 2025].
4. J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, May–June 2007.
5. The Pandas Development Team, “Pandas: Powerful Python Data Analysis Toolkit,” [Online]. Available: https://pandas.pydata.org/. [Accessed: Apr. 9, 2025].
6. NumPy Developers, “NumPy,” [Online]. Available: https://numpy.org/. [Accessed: Apr. 9, 2025].
7. SciPy Developers, “SciPy: Scientific Computing Tools for Python,” [Online]. Available: https://scipy.org/. [Accessed: Apr. 9, 2025].
8. Scikit-learn Developers, “scikit-learn: Machine Learning in Python,” [Online]. Available: https://scikit-learn.org/. [Accessed: Apr. 9, 2025].
9. Streamlit Inc., “Streamlit: The fastest way to build data apps,” [Online]. Available: https://streamlit.io/. [Accessed: Apr. 9, 2025].
10. Plotly Technologies Inc., “Plotly: The Frontend for ML and Data Science,” [Online]. Available: https://plotly.com/python/. [Accessed: Apr. 9, 2025].
11. statsmodels Developers, “statsmodels: Statistical Modeling in Python,” [Online]. Available: https://[www.statsmodels.org/.](http://www.statsmodels.org/) [Accessed: Apr. 9, 2025].
12. J. Brownlee, “How to Calculate and Interpret a Correlation Coefficient,” Machine Learning Mastery, [Online]. Available: https://machinelearningmastery.com/how-to-use- correlation-to-understand-relationships/. [Accessed: Apr. 9, 2025].
13. S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, UK: Packt Publishing, 2019.

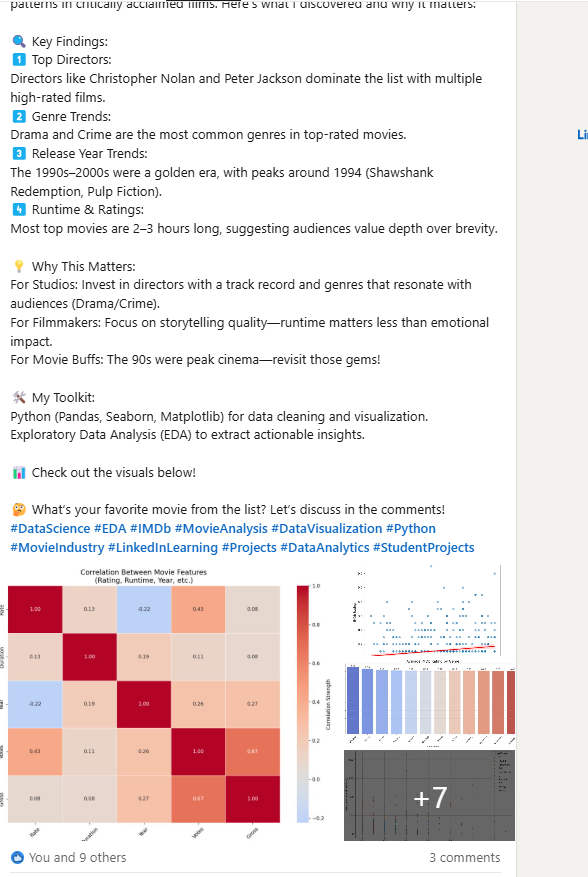
**Github Pic**

I

****

**LinkedIn Pic**

****

****