Name: MUHAMMAD ROFIAT OPEMIPO.

Title: Stage_One (Movie Analysis Python Project).

Date: 21st of October 2025.

1.0 Introduction

The dataset used for this analysis was obtained from the dataset provided (MovieLens Dataset). It contains information about movies, users, tags, and their ratings, including movie titles, genres, release years, and user IDs.

This analysis is carried out to explore patterns and trends in movie ratings and understand how factors like genre, year, and user activity influence ratings or movie recommendations.

In preparation for the analysis, the following steps were performed to prepare the data:

- Load Dataset (to import the dataset required from the zip file)
- Merged Dataset (joining the csv.files of the dataset to make it useful for analysis)
- Cleaned the Data
    1. Removed duplicates and missing values.
    2. Extracted the release year from the movie title using string operations.
    3. Converted date columns (e.g., timestamp) to datetime format for time-based analysis.
- Transform Columns
    1. Separated the movie year from title.

2.0 Features Used and Importance

Features were created from the dataset to use in data exploratory Analysis to analyze and provide meaningful insights. Which are the following:

- Release year from title:

Description: *Extracted the movie's release year from its title.*

Importance*: This helps in understanding how ratings vary by time*

- Number of genres per movie:

Description: *Counted how many genres each movie belongs to*

Importance: *It helps measure the diversity of a movie's content.*

- Primary genre of movies

Description: *Extracted the first genre listed for each movie as its main or primary genre.*

Importance: *Some recommendation models perform better when focusing on a movie's dominant category.*

- Number of tags per movie

Description: *Calculated how many tags are associated with each movie in the tags.csv file.*

Importance: *Tags capture user-generated insights about movies*

- Movie Age When Rated

Description: *Calculated the number of years between the movie's release year and the rating timestamp.*

Importance:
It helps analyze how movie age affects ratings

- Movie Rating Counts

Description: *Counted how many times each movie was rated by users.*

Importance: *This reflects popularity.*

3.0 Key Insights and Visualizations

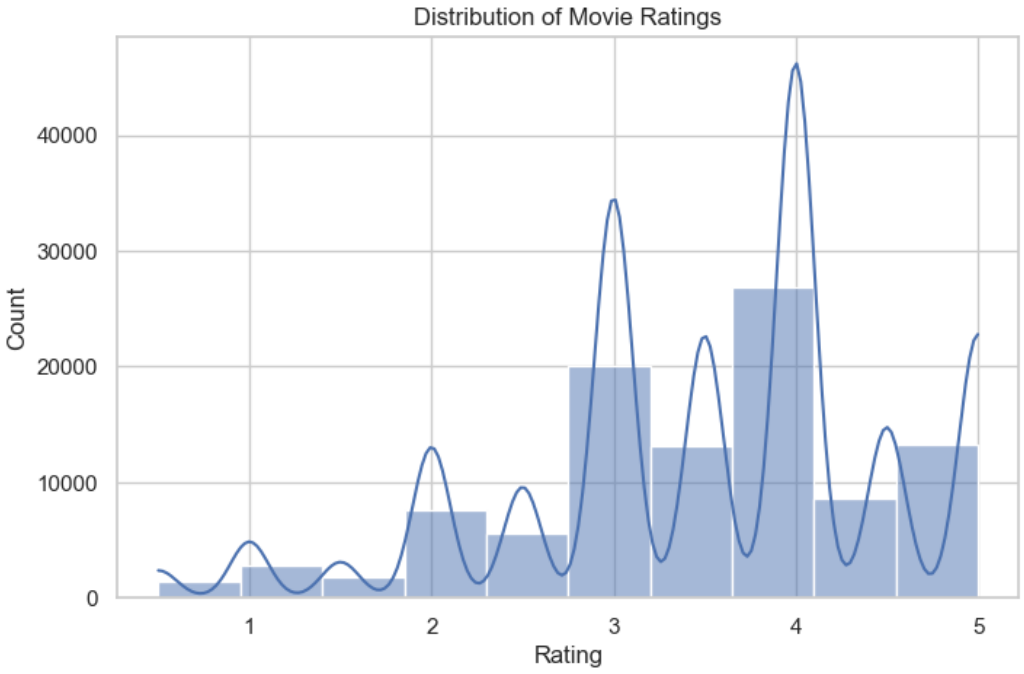

Fig 1: Movie Ratings Distribution

Insight: Most users tend to rate movies between 3.0 and 4.0, showing a bias toward positive ratings.

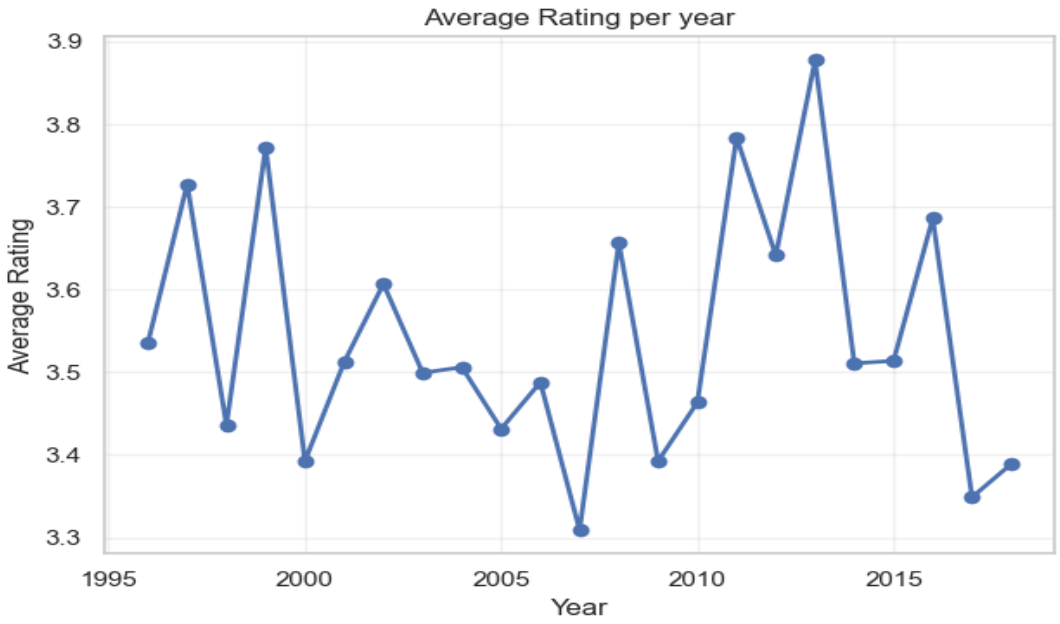

Fig 2: Average Ratings per year.

Insight: The lowest ratings were observed between 2005 and 2010. While the highest ratings were observed between 2010 to 2016

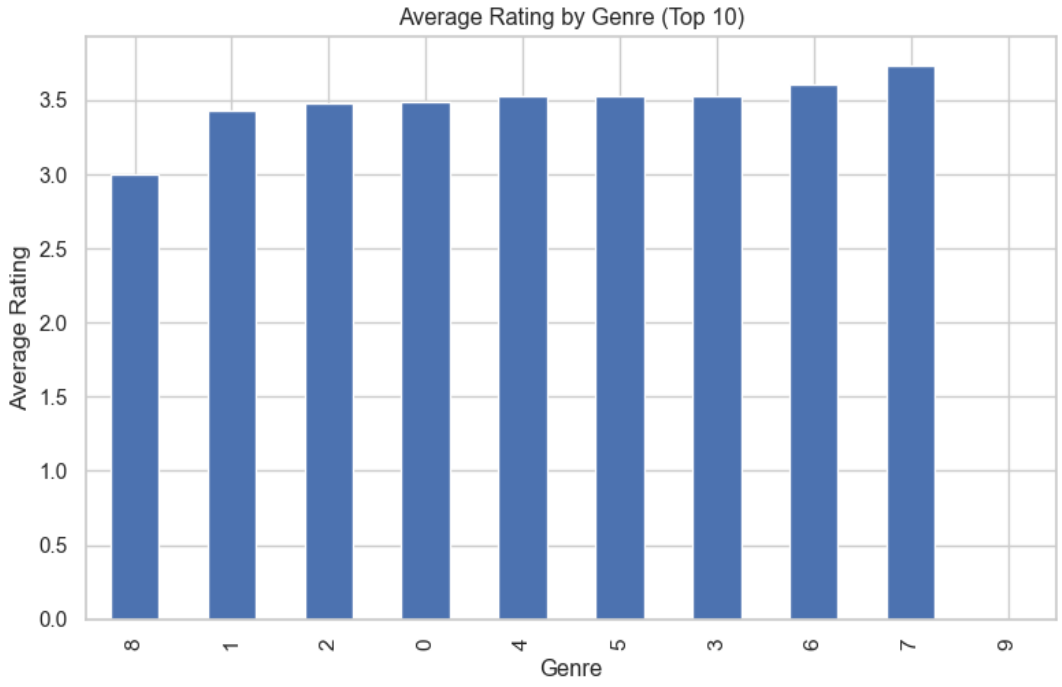

Fig 3: Average Ratings by genre.

Insight: Most genres have average ratings between 3.0 and 3.8, showing that users generally rate movies positively. Genres 6 and 7 have the highest average ratings while Genre 8 has the lowest rating, suggesting viewers are less satisfied with movies in that genre.

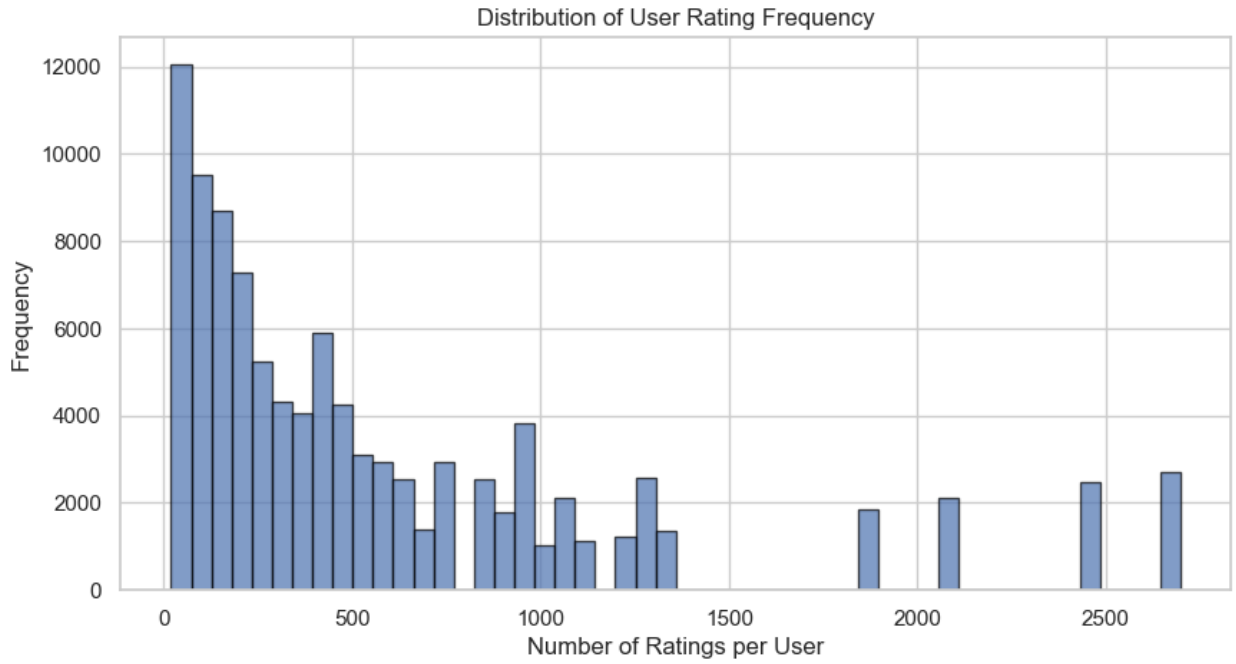

Fig 4: Distribution of User Rating Frequency.

Insight: Most users rated only a few movies, while a small number of users gave many ratings. This shows that the dataset has many non-active users and a few most active users.

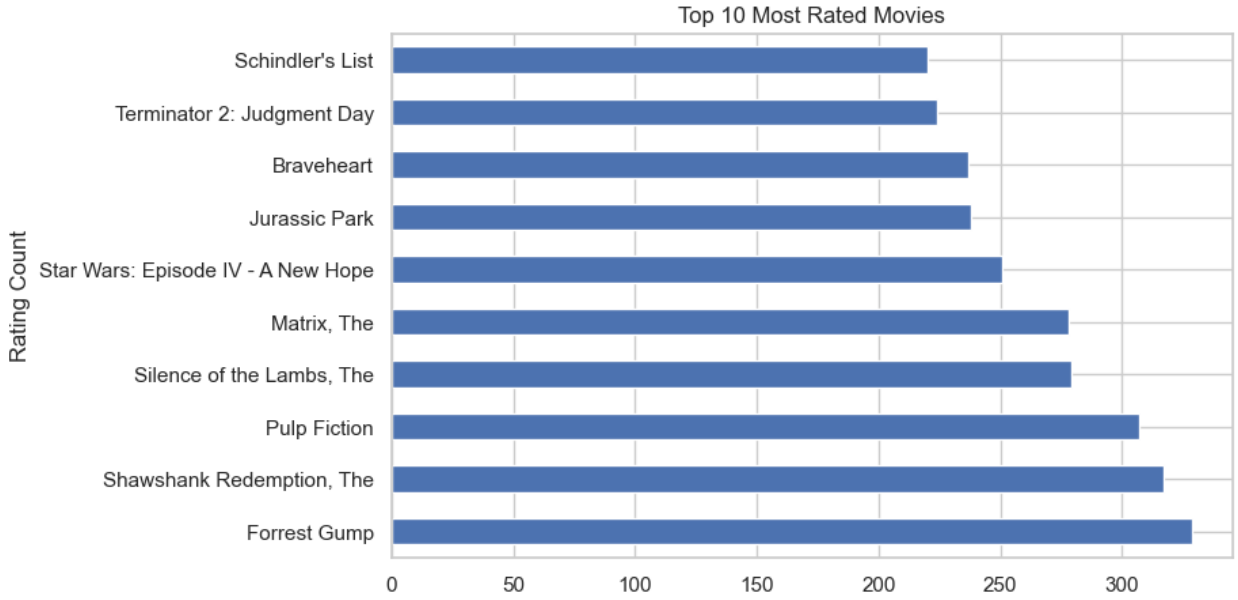

Fig 5: Top 10 Most Rated Movies.

Insight: Movies like Forrest Gump, The Shawshank Redemption, and Pulp Fiction received the most ratings. These films are popular and at the same time widely reviewed.

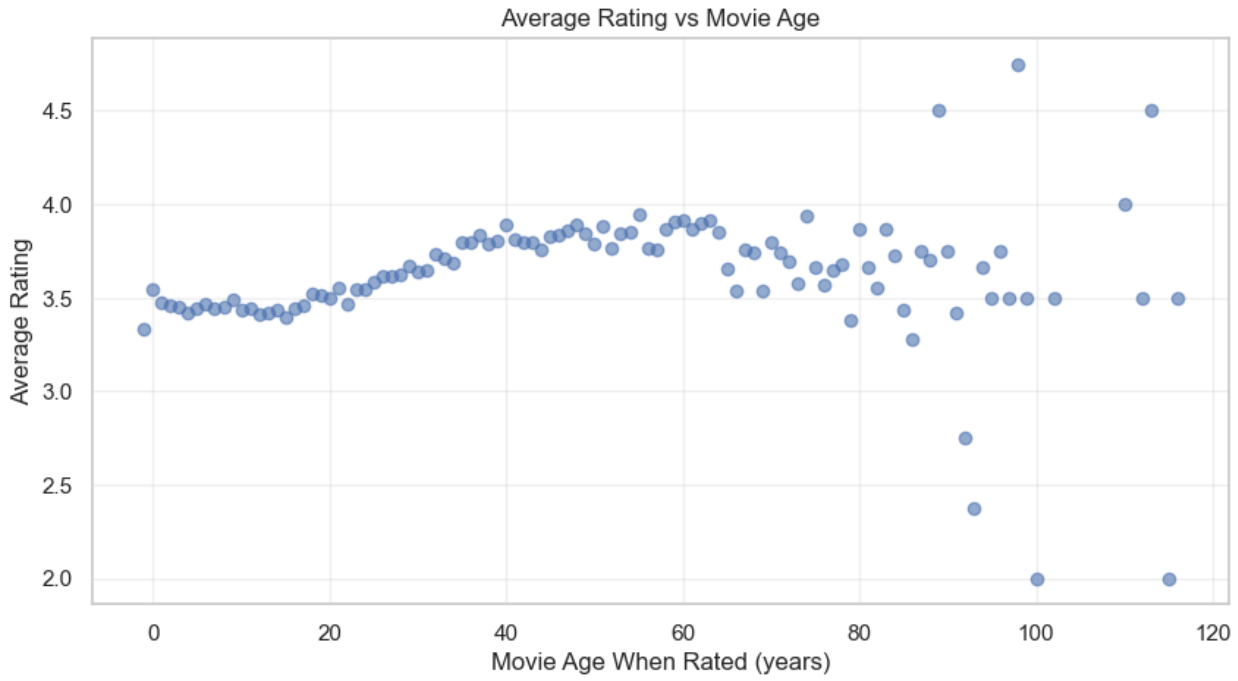

Fig 6: Top 10 Average Ratings vs Movie Age.

Insight: Movies aged between 20 to 60 years tend to have slightly higher average ratings, mostly between 3.5 and 4.0. However, very old movies (above 80 years) show that some are rated very low, while others very high.

4.0 Future Recommendations from Key Insights

- User Rating Frequency Distribution: How active users rate a lot of movies, while others rate just a few.
  Recommendations support: Most active users help improve collaborative filtering because their behavior overlaps with many others. This helps find patterns between similar users. And for less active users (cold start users), the system can rely more on movie-level features like genre and average rating.

- Most Rated Movies: Which movies have been rated the most (popular movies).
  Recommendations support: Popular movies can serve as baseline recommendations for new users who have no history yet. They can also be used to measure the accuracy of the recommendation model since they have enough ratings for comparison.

- Distribution of Movie Ratings: How users generally rate movies whether they rate high, low, or medium.
  Recommendations support: Understanding rating behavior helps in normalizing data to check if recommendations are not overly skewed toward movies with average ratings.

- Average Rating Over the Years: How movie ratings change across release years. Does older movies get higher ratings, or newer ones get more attention.
  Recommendations support: It helps in identifying time-based preferences.

- Average Rating per Genre: Which genres receive the best or worst ratings on average.
  Recommendations support: Genres are one of the strongest indicators of user taste. The model can prioritize similar high-rate genres. It also helps with content-based filtering to recommended movies with similar content (genre, storyline).

5.0 Conclusion

We observed how ratings vary across time, genres, and users. These insights are crucial for building more personalized and balanced recommendation systems that consider both user behavior and movie characteristics.