Findings Report: Movie Ratings ETL Pipeline

By Shane Henrikson

Project Overview

The "Movie Ratings ETL Pipeline" project was executed in Google Colab with the goal of building an ETL pipeline to process movie rating data. The pipeline extracted data from CSV files, performed data cleaning and transformations, and visualized key insights. This report outlines the observations, cleaning steps, and transformations achieved, along with key findings from the analysis.

Key Observations: Data Extraction

1. **Links Data:**
    - Columns: movieId, imdbId, and tmdbId.
    - Issue: 8 missing values in the tmdbId column (9734 out of 9742 rows are non-null).
2. **Movies Data:**
    - Columns: movieId, title, and genres.
    - No missing values detected.
3. **Ratings Data:**
    - Columns: userId, movieId, rating, and timestamp.
    - No missing values detected.
4. **Tags Data:**
    - Columns: userId, movieId, tag, and timestamp.
    - No missing values detected.

Key Observations: Data Cleaning

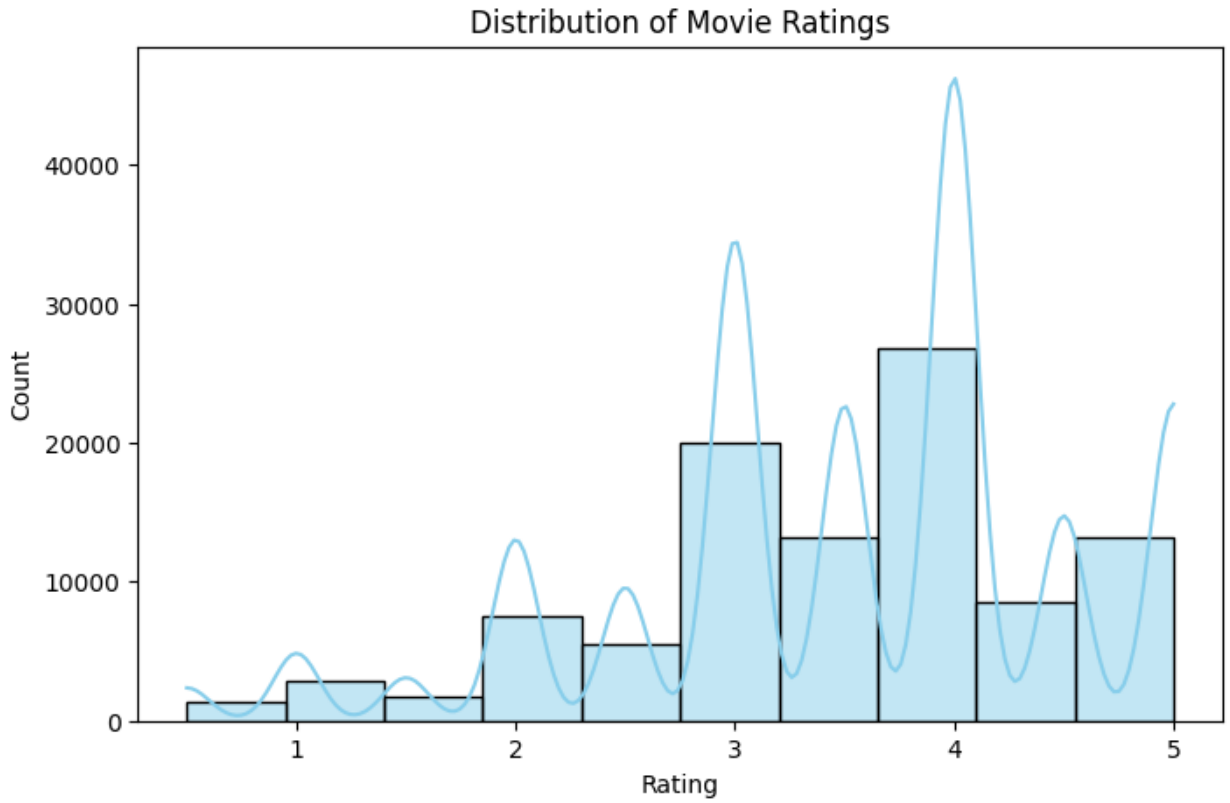
1. **Handling Missing Values:**
    - The 8 rows with missing tmdbId in links.csv were dropped, reducing the row count from 9742 to 9734.
2. **Updating Data Types:**
    - The tmdbId column in links_cleaned was converted to an integer for consistency.
    - The timestamp columns in ratings and tags were converted to readable datetime formats.

Achievements from Data Transformation

1. **Merged Links and Movies:**
    - Combined links_cleaned with movies to enrich the dataset with title and genres information.
2. **Extracted Year:**
    - Added a year column to the movies DataFrame by parsing the release year from the movie titles. This enabled year-based analysis.
3. **Filtered High Ratings:**
    - Isolated ratings ≥ 4.0, resulting in 48,580 rows. This filtering highlights highly-rated movies for further analysis.
4. **Joined High Ratings with Movie Titles:**
    - Enriched the high_ratings data with title and genres, making it more meaningful for analysis and visualization.

Graphical Findings

The analysis yielded several key visual insights for **Distribution of Movie Ratings**:



1. Overall Shape:

- The distribution is **skewed to the right**, indicating that a majority of movies tend to receive higher ratings. This is evident from the longer tail extending towards the higher rating values (4 and 5).

2. Peaks:

- There are **multiple peaks** in the distribution. This suggests that certain rating values are more common than others.
- The most prominent peak appears around a rating of **4.0**, implying that many movies receive ratings close to this value.

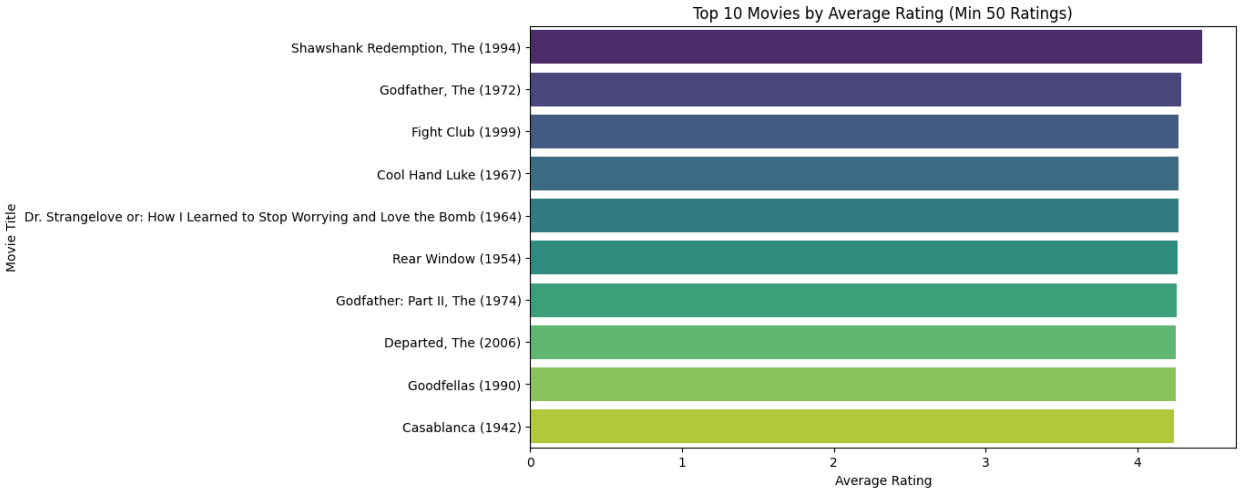
3. Frequency:

- The frequency of ratings **decreases** as the rating value increases. This means that there are fewer movies with very high ratings (close to 5) compared to those with lower or average ratings.

Interpretation:

- The right-skewed distribution and multiple peaks suggest that movie ratings are not uniformly distributed.
- The concentration of ratings around 4.0 indicates that viewers tend to favor movies that they consider "good" or "very good" rather than "excellent."
- The decreasing frequency as ratings increase suggests that achieving very high ratings is less common.

The graphical findings from the **Top 10 Movies by Average Rating (Min 50 Ratings)** bar chart:



1. Top-Rated Movies:

- The top 10 highest-rated movies, as indicated by their average rating, are as follows:
    1. Shawshank Redemption, The (1994)
    2. Godfather, The (1972)
    3. Fight Club (1999)
    4. Cool Hand Luke (1967)
    5. Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)
    6. Rear Window (1954)
    7. Godfather: Part II, The (1974)
    8. Departed, The (2006)

9. Goodfellas (1990)
10. Casablanca (1942)

2. Rating Range:

- The average ratings of the top 10 movies range from approximately 4.2 to 4.6. This indicates that the top-rated movies are generally highly regarded by viewers, with scores consistently above 4.

3. Distribution:

- The ratings are relatively evenly distributed among the top 10 movies, with no single movie dominating the list. This suggests that there are multiple films considered to be among the best, each with its own unique strengths and appeal.

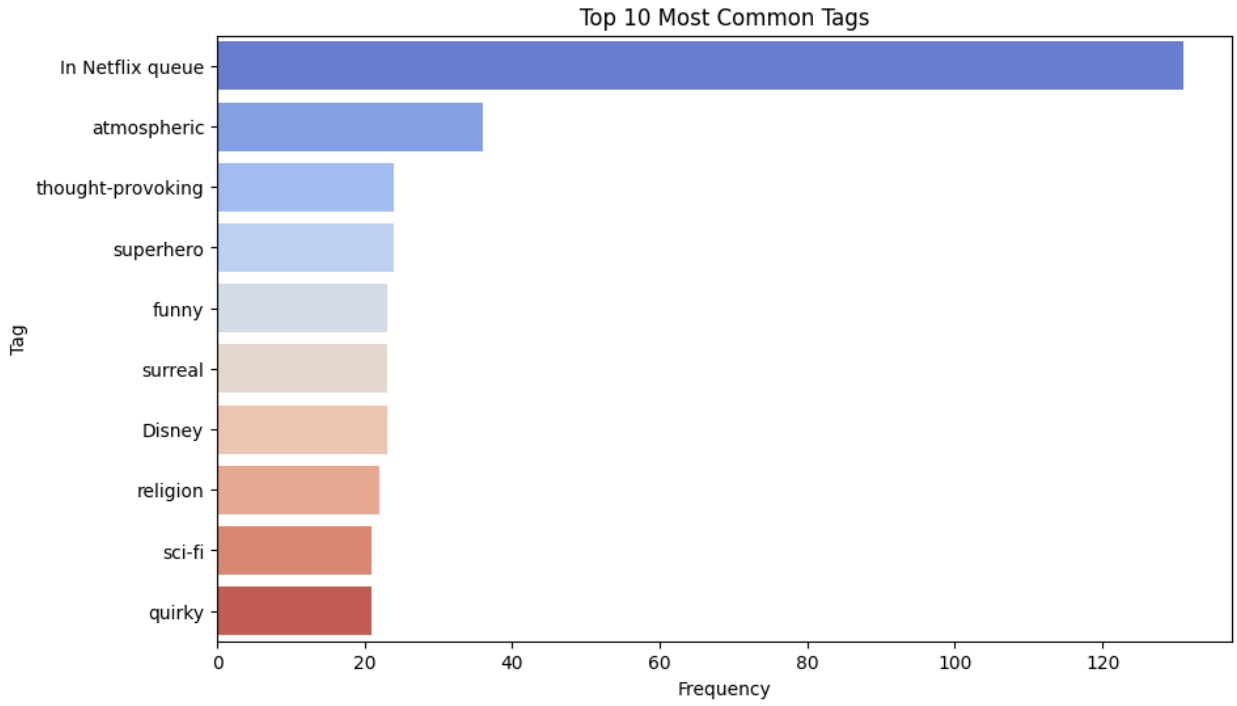
4. Notable Trends:

- The list includes both classic films and more recent releases, demonstrating that high ratings are not limited to older movies.
- Several films on the list are part of a series, such as The Godfather and The Departed, highlighting the enduring popularity of certain franchises.

Interpretation:

- The bar chart reveals that there are many excellent films available, with a diverse range of genres and release dates.
- Viewers have discerning tastes, as evidenced by the high average ratings and the spread of scores among the top 10 movies.
- The presence of both classic and modern films suggests that timeless quality is valued by viewers, regardless of the era in which the movie was made.

The graphical findings from the **Top 10 Most Common Tags** bar chart



1. Top Tags:

- The top 10 most common tags, as indicated by their frequency, are as follows:
    1. In Netflix queue
    2. Atmospheric
    3. Thought-provoking
    4. Superhero
    5. Funny
    6. Surreal
    7. Disney
    8. Religion
    9. Sci-fi
    10. Quirky

2. Frequency Range:

- The frequencies of the top 10 tags range from approximately 20 to 120. This indicates that some tags are significantly more common than others.

3. Distribution:

- The frequencies are relatively evenly distributed among the top 10 tags, with no single tag dominating the list. This suggests that a variety of tags are popular among viewers.

4. Notable Trends:

- The list includes both general and specific tags, such as In Netflix queue and superhero. This demonstrates the diversity of viewer interests.

Interpretation:

- The bar chart reveals that viewers have a wide range of preferences when it comes to tags.
- The top 10 tags are likely to be popular among many viewers, and they can be used to identify potential audience segments.
- The even distribution of frequencies suggests that there is no single dominant tag, indicating that viewers have diverse interests.