SIS: Data Collection & Preparation, Team EVOL
Worldwide Movie Gross VS. Movie Rating

Yergazy Adil 22B22B1519 & Yesserkey Dana 23B030349

Objective

Build a full pipeline: web scraping + API, cleaning, merging, basic EDA, and visualization - to study how worldwide box-office revenues relate to IMDb ratings and genres.

Data Sources

- **Web scraping**: Box Office Mojo "Top Lifetime Gross - Worldwide", 5 pages * 200 films = 1000 entries.
- **API**: OMDb API (Title, Year, imdbRating, Runtime, Genre).
- **Politeness**: User-Agent to act like a browser, api: short delays (0.1s), stop on >30 consecutive empty API responses, API results downloaded to api_results.csv.

Methodology

1. Scraping with **requests** + **BeautifulSoup**: parse the HTML table, extract Title, Lifetime Gross, Year; paginate via offset.
2. API queries to **OMDb** per (Title, Year), JSON parsed into rows, lightweight backoff; download to CSV.
3. Reproducibility: notebook can reload df_api from api_results.csv to avoid re-using the API and exhaust API keys (they are very limited).

Data Preparation and Join

Title normalization: replace long dash with short hyphen; strip spaces.
Types: "Year" - Int64; imdbRating - Float64; Lifetime Gross - Int64 (remove $ and commas).
Duration: extract minutes from Runtime via regex (\d+).
MainGenre: take the first comma-separated "main genre" from Genre.
Join: left merge on [Title, Year], left = webscraped Box Office Mojo; right = OMDb api data.
Post-clean: drop rows with any NaN, drop duplicates on [Title, Year], reset jumpin indexes.
Final columns: [Title, Year, LifetimeGross, imdbRating, Duration, MainGenre].

EDA

- IMDb ratings: min, max, mean, standard deviation computed.
- Durations: mean duration; identified shortest and longest film.
- Genre profile: value counts of main genres, most frequent genre.
- By year: yearly sum of lifetime gross, top-15 years by total gross listed.

- Main thing: association “gross - rating”: Pearson correlation between log10(LifetimeGross) and imdbRating, weak positive **r = 0.25**.
- Quantiles: lifetime gross split into five quantiles (Q1-Q5), mean raiting by quantile reported.

Visualizations

1. Histogram of IMDb ratings (20 bins).
2. Bar chart of total worldwide gross by year.
3. Horizontal bar chart of genre counts (MainGenre).
4. Scatter: imdbRating vs log10(LifetimeGross) with linear regression line; caption includes Pearson r ≈ 0.25.

Key Findings

- Ratings cluster around the mean with fewer extremes (6.2-7.2)
- Genre distribution is skewed with several dominant categories (where Action is the king).
- There was a huge drop in gross due to COVID-19, after which the film industry did not recover much, or they are simply making bad films :)
- The gross-rating link is weakly positive (r ≈ 0.25): we proved that higher-grossing films tend to have **slightly** higher ratings.

Limitations

- OMDb API is strictly limited, we had to make several keys and use it very carefully. But after a couple runs we thought of an idea to save all the data once and forever.
- Matching by (Title, Year) is sensitive to naming variants, despite normalization. We could have made the normalization even stricter: e.g., we could have make everything in lower index, remove all spaces and numbers, etc. However, after we did what we did, and saw that we had lost very few records, we decided to leave it as it was.
- Dropping rows with any missing values can bias results toward films well covered by OMDb. Like, the movie could have been unpopular and somehow incorrectly listed in the database. But considering that we took the most profitable ones, that option was out of the question anyway.
- OMDb aggregates community and curated inputs; field accuracy (Runtime, Genre) may vary.
- Correlation is not causation; franchise effects, release windows, and marketing are unobserved inaccuracies.