Google Data Analytics Capstone Project

Project Name: Genre-Based Analysis of IMDb Ratings to Online Streaming Platform Data-Driven Decisions

Prepared by: Prawit Pongpipat

Scenario

You are a junior data analyst working for a business intelligence consultant. You have been at your job for six months, and your boss feels you are ready for more responsibility. He has asked you to lead a project for a brand-new client — this will involve everything from defining the business task all the way through presenting your data-driven recommendations. You will choose the topic, ask the right questions, identify a fresh dataset and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

Step 1: Ask

1. What type of company does your client represent, and what are they asking you to accomplish?

   An online streaming platform, which they want to make data-driven decision about **which movie genres are most likely to receive highly positive reviews** for the upcoming exclusive movie on their platform.

2. What are the key factors involved in the business task you are investigating?
   - Genre of the movie
   - Year of release
   - Popularity
   - User rating
3. What type of data will be appropriate for your analysis?
   - Top-rated movies across multiple years
   - Metadata (Genre, year, rating, etc.)
4. Where will you obtain that data?

   Public online domain datasets such as Kaggle, Medium.

5. Who is your audience, and what materials will help you present to them effectively?
   - Online streaming platform’s content strategy team
   - Directors and producers

Business Task

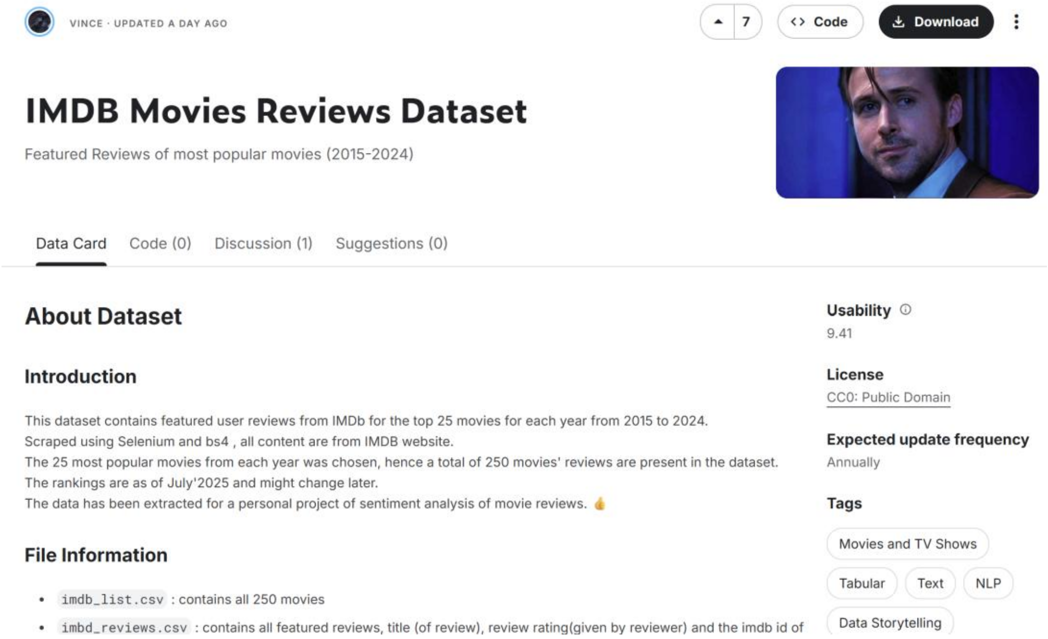
Identify which movie genres will receive the most positive reviews, to guide the online streaming platforms and the director's data-driven decision on what genre of exclusive movies to pursue.

Step 2: Prepare

1. Where is your data located?

This dataset is publicly located on Kaggle: "Popular Movies IMDB Reviews Dataset" by user VINCE.

https://www.kaggle.com/datasets/shivvm/popular-movies-imdb-reviews-dataset



IMDB Movies Reviews Dataset

Featured Reviews of most popular movies (2015-2024)

About Dataset

Introduction

This dataset contains featured user reviews from IMDb for the top 25 movies for each year from 2015 to 2024.
Scraped using Selenium and bs4 , all content are from IMDB website.
The 25 most popular movies from each year was chosen, hence a total of 250 movies' reviews are present in the dataset.
The rankings are as of July'2025 and might change later.
The data has been extracted for a personal project of sentiment analysis of movie reviews. 👍

File Information

- `imdb_list.csv` : contains all 250 movies
- `imbd_reviews.csv` : contains all featured reviews, title (of review), review rating(given by reviewer) and the imdb id of

Usability
9.41

License
CC0: Public Domain

Expected update frequency
Annually

Tags
Movies and TV Shows
Tabular
Text
NLP
Data Storytelling

2. How is the data organized?

In imdb_list.csv, the dataset is organized by the author as follows:

index	id	title	rating	genre	year

3. Are there issues with bias or credibility in this data? Does your data **ROCCC**?
    - ✅ **Reliable:** Data consistently formatted and sourced from IMDb, a credible movie rating platform.
    - ✅ **Original:** Dataset was collected and arranged by the dataset author, which is extracted from his personal project of sentiment analysis of movie reviews.
    - ✅ **Comprehensive:** Includes a wide range of fields such as ratings, genres, reviews, and year of release.
    - ✅ **Current:** Dataset coverage from 2015 to 2024 (As of July 2025).
    - ✅ **Cited:** Dataset is hosted on a reputable platform with clear licensing.
4. How are you addressing licensing, privacy, security, and accessibility?
    - **Licensing:** The dataset is publicly accessible on Kaggle as CC0: Public Domain.
    - **Privacy:** No sensitive or personal information appeared.
    - **Security:** Stored online on a secure and public platform, Kaggle.
    - **Accessibility:** Data is in .csv format, which is readable by Excel, Google Sheets, etc.
5. How did you verify the data's integrity?
    - No missing essential fields.
    - Data types are consistent.
    - Manually reviewed to ensure expected formats.
6. How does it help you answer your question?

    With this dataset, I can identify the relationship between genre and IMDb rating trends among data and find out which genres receive the most positive feedback. These insights will directly inform the online streaming platform's decision on what genre to pursue for their upcoming exclusive movies.

7. Are there any problems with the data?
    - Some genres are combined as one field ("Action, Adventure" vs "Action"), which require normalization (First Normal Form - 1NF).
    - Some fields (such as review title, review) in imdb_review.csv may be irrelevant for the core business task and will be excluded from analyzing.
    - There are two duplicate ids in imdb_list.csv (Deadpool and Deadpool & Wolverine), which require editing.

Step 3: Process

1. What tools are you choosing and why?

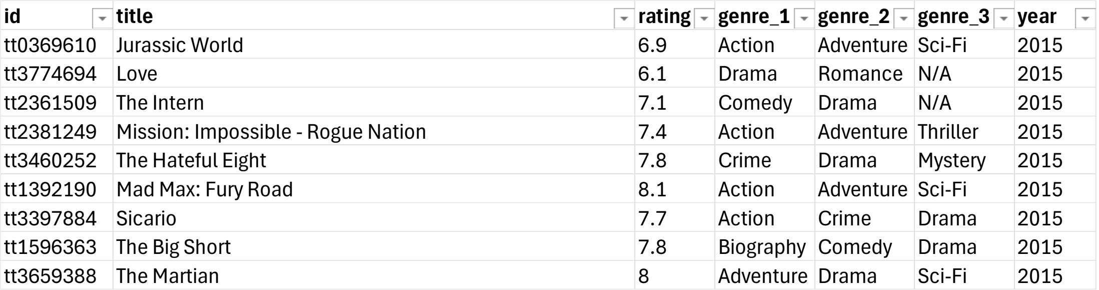
   I choose spreadsheet (Microsoft Excel) for data analysis because this dataset contains only 250 rows, which is manageable to work with in a spreadsheet environment.

2. Have you ensured your data's integrity?

   Yes, because this dataset is from **Kaggle** and released under a **CC0: Public Domain license**, which ensures it is freely reusable and trustworthy.

3. What steps have you taken to ensure that your data is clean?
   - Skimmed the dataset manually at the first round.
   - Checked for missing or blank values in key columns (By adding N/A).
   - Normalized **genre** where multiple genres are listed in a single cell.
4. How can you verify that your data is clean and ready to analyse?
   - All required fields are populated.
   - No formatting errors or data type issues remain.
   - The dataset is structured and no error, allowing for aggregation.
   - Charts and summary calculations run without errors.
5. Have you documented your cleaning process so you can review and share those results?

   Yes, and this is an example of my spreadsheet I had done.



id	title	rating	genre_1	genre_2	genre_3	year
tt0369610	Jurassic World	6.9	Action	Adventure	Sci-Fi	2015
tt3774694	Love	6.1	Drama	Romance	N/A	2015
tt2361509	The Intern	7.1	Comedy	Drama	N/A	2015
tt2381249	Mission: Impossible - Rogue Nation	7.4	Action	Adventure	Thriller	2015
tt3460252	The Hateful Eight	7.8	Crime	Drama	Mystery	2015
tt1392190	Mad Max: Fury Road	8.1	Action	Adventure	Sci-Fi	2015
tt3397884	Sicario	7.7	Action	Crime	Drama	2015
tt1596363	The Big Short	7.8	Biography	Comedy	Drama	2015
tt3659388	The Martian	8	Adventure	Drama	Sci-Fi	2015

Step 4: Analyze

1. How should you organize your data to perform analysis on it?
   - Ensuring each column is clearly labeled.
   - Splitting multi-genre entries into separate columns for better filtering.
   - Creating summary tables using pivot tables to group data by genre and average IMDb ratings.

2. Has your data been properly formatted?

   Yes. I ensured all ratings are in number format (e.g., 8.7, 6) and genres are formatted consistently.

3. What surprises did you discover in the data?

   I found out that the average IMDb ratings between 2020 and 2022 were significantly lower compared to other years (2015 to 2019, 2023 and 2024). And this period aligns with the COVID-19 pandemic, which may have impacted on production quality, storytelling depth, or audience expectations.

4. What trends or relationships did you find in the data?
   - **Animation and Biography** had the highest average IMDb ratings.
   - **Action** had the greatest number of movies.
   - **Horror** had the lowest average IMDb ratings.
   - **12th Fail** had the highest IMDb rating at 8.7, whereas **Justice League** had the lowest IMDb rating at 6.0.
5. How will these insights help answer your business questions?

   These findings support the business task of selecting a genre for the online streaming platform's next exclusive production. By identifying genres such as Animation or Biography that consistently receive higher ratings and more positive reviews.

   With this, the online streaming platform can make a data-driven decision on where to invest. Moreover, this analysis also helps balance rating and potential popularity to align with audience expectations.

Step 5: Share

1. Were you able to answer the business question?

   Yes, by analyzing the relationship between genres and IMDb ratings, I identified which genres tend to receive the most positive rating. This insight can help the online streaming platform make a data-driven decision on selecting a genre for their next exclusive production.

2. What story does your data tell?

   This dataset reveals key insights about audience. **Animation and Biography** genres had the highest average IMDb ratings, which suggesting strong viewer appreciation. In the other hand, **Horror** films consistently received the lowest average IMDb ratings, showing weaker reception. While **Action** was the most frequently occurring genre in the dataset, it did not perform as well in terms of average ratings.

Furthermore, the movie **“12th Fail”** achieved the highest IMDb rating at 8.7, whereas “**Justice League”** had the lowest rating at 6.0. These findings highlight a gap between quantity and quality across genres and help identify which types of genres or content are more favourably received by audiences.

3. How do your findings relate to your original question?

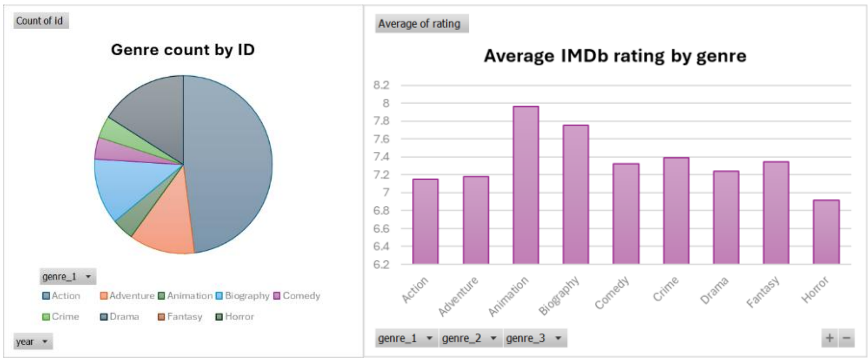
   The original question asked **which movie genres are most likely to receive highly positive reviews**. The findings relate to this by highlighting the genres that are historically best rated. This gives actionable insights to prioritize genres like Animation or Biography, which are more likely to succeed with audiences.

4. Who is your audience? What is the best way to communicate with them?

   The content strategy team and **creative directors** at the online streaming platform. And the best way to communicate with them is through a **concise slide presentation** that combines clear, high-level insights, and practical recommendations.

5. Can data visualization help you share your findings?

   Yes, data visualization is a key tool in communicating insights clearly and persuasively. These are my examples of charts:



Step 6: Act

1. What is your final conclusion based on your analysis?

   These insights suggest that the online streaming platform should consider producing content in **Animation** or **Biography** genres to maximize viewer satisfaction and positive reviews.

2. How could your team and business apply your insights?

   The content strategy team can use these insights to prioritize genres with proven audience approval for their upcoming exclusive productions. This reduces risk and increases the chance of releasing a successful and profitable movie.

3. What next steps would you or your stakeholders take based on your findings?
   - Shortlist potential concepts in the Animation or Biography genres.
   - Conduct further analysis on more specific audience segments.
   - Begin market testing to validate interest in the chosen genre.
   - Consider analyzing competitors or recent performance for added context.
4. Is there additional data you could use to expand on your findings?
   - Additional demographic data to better understand audience segments.
   - Online streaming viewership data to measure performance alongside ratings.
   - Sentiment analysis on user reviews for deeper qualitative insights.
   - Production details to analyze what factors contribute to high ratings within top genres.
