Emma Luk
Anna Fariha
CS 6959
April 1st, 2025

Data Profiling IMDb

The IMDb database has 6 tables: actors, casts, directors, genre, movie, movie_directors.

The actor table has 4 columns: id, fname, lname, gender.

- There are 3442863 unique ids, ie 3442863 unique actors.
- There are 297967 unique first names.
  - The most common first names(top 10) are: John, David, Michael, Robert, Paul, Anna, Peter, Sarah, Jennifer, Maria
  - The least common first names(bottom 10) are: Rohay, Scrapey P., Jelovus Steve, Aimane, Tawquir, Hong-ryeol, Roldolfo, Ovidel, Kareem J., Lachauncey
- There are 612647 unique last names.
  - The most common last names(top 10) are: <empty>, Smith, Williams, Taylor, Wilson, Lee, Thomas, Jones, Johnson, Brown
  - The least common last names(bottom 10) are: Drusche, Iversen Gangvik, Desremaux, Kacian, Escarpeta, Bissila, Estelle-Vaisset, Dresdner Philharmoniker, Machada, Filgis
- There are 2 genders recorded: F(female), M(male).
  - The ratio of F to M actors is 1903167 to 1539696, respectively.

The casts table has 3 columns: pid, mid, role.

- Pid is a foreign key to the actors' table id.
- Mid is a foreign key to the movie's table id.

The directors table has 3 columns: id, fname, lname.

- There are 583174 unique ids, ie 583174 unique directors.
- There are 72107 unique first names.
  - The most common first names(top 10) are: David, Michael, John, Peter, Paul, Daniel, Robert, Chris, Mark, James.
  - The least common first names(bottom 10) are: Merran, Numael, <empty>, Chang Soo, Lliph Amen, Djoeke, Reef, Hiawitha, Nalan, Valmir.
- There are 163236 unique last names.
  - The most common last names(top 10) are: <empty>, Smith, Lee, Williams, Johnson, Jones, Brown, Kim, Miller, Davis.
  - The least common last names(bottom 10) are: Beauchesne-Rondeau, Bassman, Ardussi, Benevich, Arcady, Bakry, Beautour, Andry, Amponsah, Albacete.

The genre table has 2 columns: mid and genre.

- Mid is a foreign key to the movie's table id.
- There are 27 unique genres: Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, Game-Show, History, Horror, Music,

Musical, Mystery, News, Reality-TV, Romance, Sci-Fi, Short, Sport, Talk-Show, Thriller, War, Western.

The movie table has 3 columns: id, name, year.

There are 2383721 unique ids, ie 2383721 unique movies.

The oldest movie is from 1519.

The newest movie is from 2025.

The movie_directors table has 2 columns: did, mid.

Did is a foreign key to the directors' table id.

Mid is a foreign key to the movie's table id.

The IMDb dataset provides a comprehensive overview of movies, actors, and directors. There are over 3.4 million unique actors and 2.3 million movies. I was surprised to see that there were more female actors (55%) than male actors. The most common first names for BOTH actors and directors are "John" and "David," while rare names such as "Rohay" and "Merran" are much less frequent. The genre table includes 27 categories. The oldest movie dates back to 1519, and the most recent is from 2025, showing a wide range of movie production over time. I was really surprised that the oldest movie was 1519; I didn't even know movies could be recorded back then.