Lan (Sherry) Lou
DSCI 510 Final Project Description

Motivation

I am passionate about the movie industry, and the motivation of this project stemmed from my desire to know the factors that might influence the popularity/rating of a certain movie. Given the top 250 movies rated by IMDB users, it would be interesting to look at how box office, directors, casts, release year, worldwide popularity have played a role in the high rating of such movies. The project result would be beneficial to movie industry investors and production companies to foresee the public feedback of a future movie.

Brief of data

With one web scraping and two usages of API, I have input all three datasets into a database called "top_250_movies.db". Each of the three datasets is imported as a table, so there are a total of three tables within this database. All three of them are linked by the primary key of "imdb_id" because it is a unique identifier of each movie. Below is a detailed description of what are included in each dataset:

1) **"top_250_basic_movies_info"** dataset: (ranking INTEGER, imdb_id TEXT, director TEXT, actor1 TEXT, actor2 TEXT, rating REAL)
2) **"more_movies_info"** dataset: (imdb_id TEXT, movie_name TEXT, year INTEGER, release_time TEXT, runtime INTEGER, genre TEXT, language TEXT, country TEXT, awards TEXT, box_office REAL)
3) **"movie_aliases"** dataset: (imdb_id TEXT, alias TEXT)

Sources

Although all three datasets are related to IMDB movie data, they come from different data sources and details are listed below.

1) **"top_250_basic_movies_info"** dataset: web scraping from IMDB ranking site https://www.imdb.com/chart/top where top 250 rated movies are listed
2) **"more_movies_info"** dataset: used the "imdb_id" extracted from each of the top 250 movies in the first dataset and use an API https://rapidapi.com/rapidapi/api/movie-database-imdb-alternative/ to get more movie information on each of the top 250 movies to complement the first dataset
3) **"movie_aliases"** dataset: used the "imdb_id" extracted from each of the top 250 movies in the first dataset and use an API https://rapidapi.com/amrelrafie/api/movies-tvshows-data-imdb/ to get all aliases of each of the top 250 movies as an indication of the popularity of movies worldwide

Analysis performed

Both descriptive and inferential analysis are performed regarding the data I extracted from above.

Descriptive analyses are mainly performed using SQL query language and include:

1) Display the top five rated movies by IMDB users
2) Top 250 movies' rating description: average rating, minimum rating, maximum rating
3) Top 250 movies' runtime description: average runtime, minimum runtime, maximum rating

4) Top five directors who have the most movies on the top 250 list
5) Top five actors who starred in the most movies on the top 250 list
6) Most popular movies around the world using number of aliases as the indication
7) The most popular three genres for the movies to be on the top 250 list
8) The most popular three movie production countries on the top 250 list
9) Top 250 movies' release year distribution graph
10) Top 250 movies' runtime distribution graph

Inferential analyses are performed using correlation graphs and linear regression model and include:

1) Correlation between rating and runtime, and linear relationship exploration
2) Correlation between rating and box office, and linear relationship exploration
3) Correlation between rating and number of aliases, and linear relationship exploration

Conclusions

Based on the analyses mentioned above, there are a few insights drawn accordingly.

Descriptive analyses results:

1) The top five rated IMDB movies are

Top 1: The Shawshank Redemption with rating of 9.220649578673955 and box office of $28699976.0
Top 2: The Godfather with rating of 9.14744669786617 and box office of $134966411.0
Top 3: The Godfather: Part II with rating of 8.980915536629867 and box office of $47834595.0
Top 4: The Dark Knight with rating of 8.97405415507201 and box office of $534858444.0
Top 5: 12 Angry Men with rating of 8.940342076396444 and box office of $N/A

2) Top 250 movies' rating description

The average rating of the top 250 movies is 8.260663574127554
The minimum rating of the top 250 movies is 8.019925624066646
The maximum rating of the top 250 movies is 9.220649578673955

3) Top 250 movies' runtime description

The average runtime of the top 250 movies is 129.684 mins.
The minimum runtime of the top 250 movies is 45 mins.
The maximum runtime of the top 250 movies is 321 mins.

4) Top five directors who have the most movies on the top 250 list

Top 1: Akira Kurosawa (dir.) with 7 movies.
Top 2: Christopher Nolan (dir.) with 7 movies.
Top 3: Martin Scorsese (dir.) with 7 movies.
Top 4: Stanley Kubrick (dir.) with 7 movies.
Top 5: Alfred Hitchcock (dir.) with 6 movies.

5) Top five actors who starred in the most movies on the top 250 list

Top 1: Robert De Niro with 6 movies.
Top 2: Tom Hanks with 5 movies.
Top 3: Leonardo DiCaprio with 5 movies.

Top 4: Charles Chaplin with 5 movies.
Top 5: Toshirô Mifune with 4 movies.

6) Most popular movies around the world using number of aliases as the indication

Top 1: Star Wars : distributed to 176 countries.
Top 2: Star Wars: Episode VI - Return of the Jedi : distributed to 125 countries.
Top 3: Star Wars: Episode V - The Empire Strikes Back : distributed to 123 countries.
Top 4: The Lord of the Rings: The Fellowship of the Ring : distributed to 102 countries.
Top 5: Harry Potter and the Deathly Hallows: Part 2 : distributed to 101 countries.

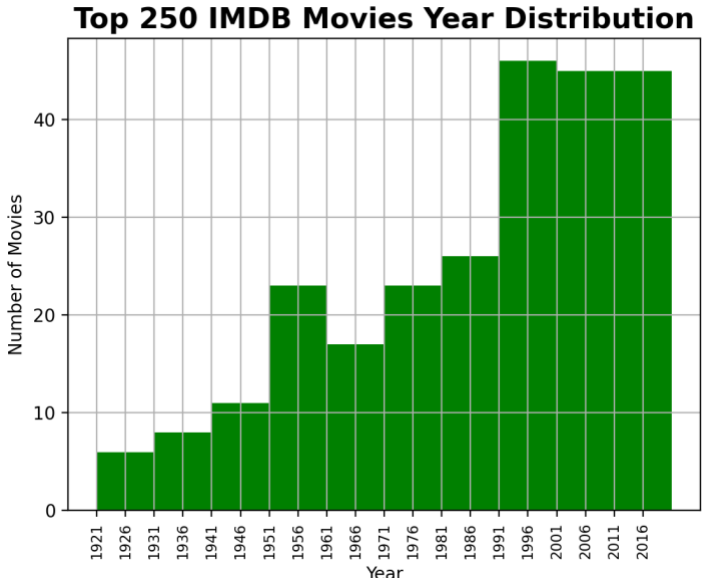
7) The most popular three genres for the movies to be on the top 250 list

Top 1: Drama with a frequency of 184
Top 2: Adventure with a frequency of 56
Top 3: Crime with a frequency of 56

8) The most popular three movie production countries on the top 250 list

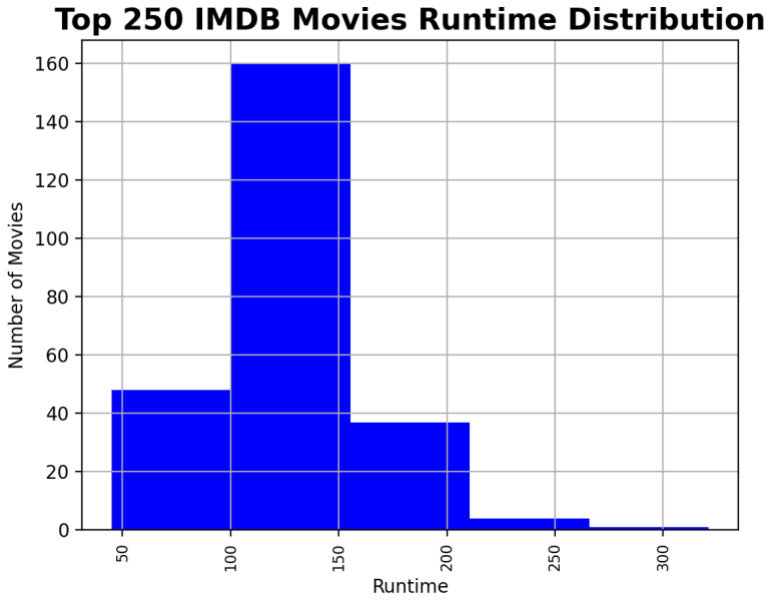
Top 1: United States with a frequency of 170
Top 2: United Kingdom with a frequency of 49
Top 3: France with a frequency of 49

9) Top 250 movies' release year distribution graph



Most top 250 movies tend to aggregate toward the release year of 1991 to 2016. However, there is a small peak during the year from 1951 to 1961.

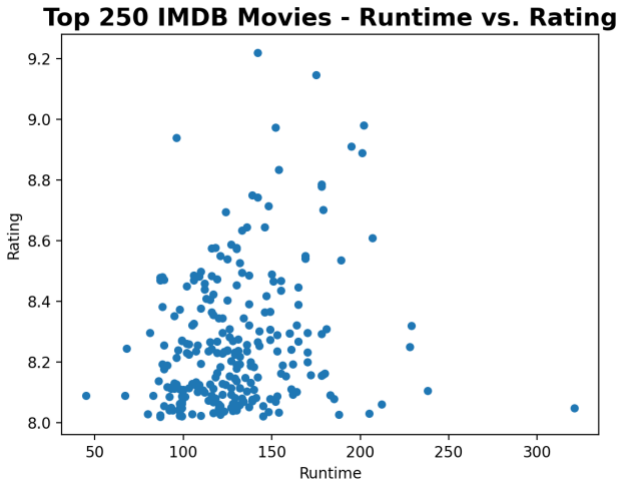
10) Top 250 movies' runtime distribution graph



160 out of 250 (64%) of the top-rated movies have the runtime from 100 minutes to 150 minutes. However, there are still movies that are very long and in the 300 minutes range.

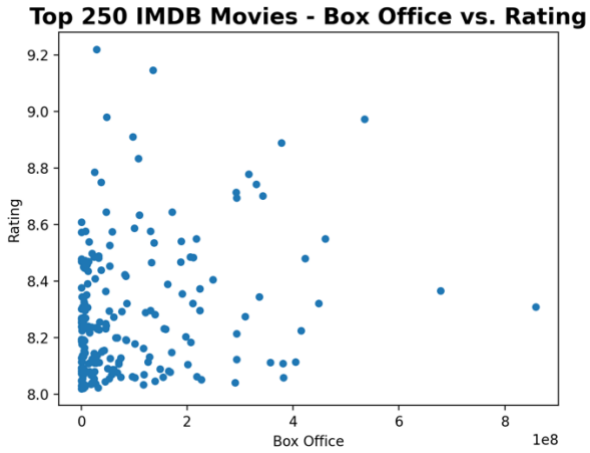
Inferential analyses results:

1) **Correlation between rating and runtime, and linear relationship exploration**



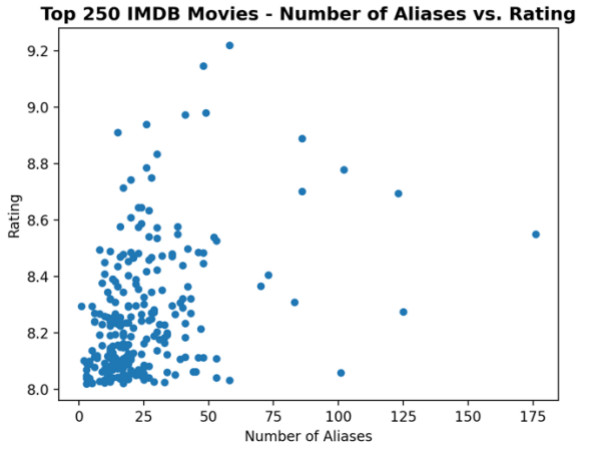
- Correlation formula: rating = 0.001473744924136164 * runtime + 8.069542437385877
- R-square: 0.04679693147877373
- Because the r square value is so low (below 0.4), the variation of the result cannot be explained well by the correlation formula calculated. Therefore, there is no/low correlation between runtime and rating.

2) **Correlation between rating and box office, and linear relationship exploration**



- Correlation formula: rating = 4.184909168070268e-10 * box office + 8.237928442858056
- R-square: 0.05713182858605123
- Because the r square value is so low (below 0.4), the variation of the result cannot be explained well by the correlation formula calculated. Therefore, there is no/low correlation between box office and rating.

3) Correlation between rating and number of aliases, and linear relationship exploration



- Correlation formula: rating = 0.003941997840749098 * number of aliases + 8.16356428331422
- R-square: 0.1351988549573893
- Because the r square value is so low (below 0.4), the variation of the result cannot be explained well by the correlation formula calculated. Therefore, there is no/low correlation between number of aliases and rating. However, the relationship between number of aliases and rating is slightly stronger than rating's relationship with runtime or box office.