Charlene Guerrero
Dr. Whalen
DSCI 101 001
11 December 2023

DSCI101 Final Project: An Analysis of Movies within the IMDb database

Introduction

For my project, I will be analyzing data from the Internet Movie Database (IMDb), an online database containing information about films, TV series, video games, and other media. Existing since 1990 this database has accumulated entries on 10,206,303 unique titles or works (as of November 2023) and contains information on their casts, production crews, regions, titles, genres, and average ratings to name a few variables. There is no charge to access and contribute to IMDb database for registered IMDb users who submit new material and suggest edits to title entries. Meanwhile, industry professionals, such as producers and actors, can pay for membership and the ability to provide their resumes and details on their productions. Notably, IMDb.com Inc., a subsidiary company of Amazon that owns and operates the database, monitors and approves new data entries and edits to ensure accurate reports of each title.

Since the IMDb website was the 52nd most visited website in 2019 and boasts 83 million registered users, this database is constantly gaining new data and has become one of the largest, most comprehensive movie databases online. With such frequent and widespread usage, this database would be interesting to analyze for casual filmgoers and professional film critics alike. Therefore, this project will focus primarily on the 658,561 movie entries within the dataset and not on information about other title types, such as TV series and video games.

Finally, on a personal note, I grew up watching movies with my family every Friday night and have developed a deep fondness for cinema. Recently, I have been trying to watch more foreign films and I enjoy watching all the films from a given director. For instance, I am currently on a Wes Anderson binge and watching all his films from his earliest release, “Bottle Rocket (1996),” to his latest, “Asteroid City (2023).” Therefore, I took this project as an opportunity to find which countries release the most films, which directors are most popular on the IMDb website, and possible reasons for why these directors are the most popular. Ultimately, this project will demonstrate top regions, directors, and genres to consider for IMDb users and myself when choosing a new film to watch for the next movie night.

Major Questions

While considering the IMDb database, I was first curious to see if different countries released more titles or movies than others and what this movie released looked like over time. Then, I was interested to observe the overall distribution of directors’ ratings. Taking these two questions together, I then decided to look at the most popular directors in the United States, the region that has released the most films since the genesis of the IMDb database. Specifically, I based popularity on the greatest number of votes received by reviewers on the IMDb website. Finally, I was curious to see if certain genres were associated with these popular directors.

To better shape the analysis in this project, I have listed the questions I will be answering below. A list of the datasets and variables used to answer these questions is also provided in the corresponding Rmarkdown file.

1. What are the top 5 regions that release the most titles (not just movies) overall? What are the top 5 regions that release the most movies overall? Which region has released the most movies in 2023?
2. Looking at each of top 5 regions that released either the most titles or movies overall, what is the trend in movie release from 1894 (year the first movie was recorded) to 2023?
3. What is the distribution director ratings (mean rating of a all a director's films) in the IMDb database?
4. Looking only at the United States (region = US), which directors (top 5) have received the most total movie rating votes? How many films has each released? What is the average film rating for each of these directors?
5. Do the top five directors from question 4 release movies associated with specific genres?
6. When Christopher Nolan releases films that are not of the drama genre, do the films rate the same? Are they rated higher or lower?

Analysis and Discussion

Since IMDb is an internationally used database, I first aimed to depict how movie release differed across regions overall. Before doing this, however, I wanted to see if overall titles, not just movies, differed across regions. To answer these two questions, I first grouped the IMDb data by region and sorted the regions from highest to lowest number of titles or movies released. Interestingly, when looking at all title types, Germany released the most with 4,428,493 titles overall. This was followed by France (4,426,116), Japan (4,423,737), India (4,369,497), and Spain (4,344,207) (**figure 1A**). In contrast, when I filtered for only movies, the United States (337,870 movies), United Kingdom (169,007), India (100,691), Japan (97,335), and Canada (96,760) released the most films (**figure 1B**). The United States also released the most movies in 2023 with 7,235 total movies released. Interestingly, while the top five regions when looking at titles released around the same number of titles (4.3-4.4 million), the United States released more than double the number of movies compared to United Kingdom, the region that released the second greatest number of movies. Furthermore, it appears that from 1894, when the first film was recorded, to 2023, each of the top regions released more titles and films with time (**figures 1C-D**). There was also a steep increase in movies released for the US and Britain starting in 2000s, and a similarly sharp increase in titles released for Germany and India starting in the 2000s. Overall, the top regions clearly differed by total titles and movies released, and this may be an interesting finding to further analyze in subsequent analyses of this database. However, moving forward for this project, we will be considering director ratings of US directors as the United States was the region with the most movies released.

Before looking at US directors, understanding the director ratings--or the average ratings of all a director’s films--across all regions in the IMDb dataset would be helpful. Since the

datasets were already quite large and taking a long time to load, even after filtering for only movies, I decided to focus on one director for each movie. Specifically, since many movies had multiple authors listed, I decided to take the first author listed for each movie by separating the directors, which were stored in a single string variable, into separate values. The distribution of director ratings followed a normal distribution with most observations falling around a movie rating score of 7 (**figure 2A**).

Next, looking only at the United States, I decided to see which directors were the most popular in terms of total movie rating votes received. I accomplished this by grouping by the directors and finding the sum of all the votes they had received. In this way, the top five directors were Christopher Nolan (67,103,427 total votes), Steven Spielberg (36,197,693), George Lucas (31,709,719), James Cameron (28,996,796), and Quentin Tarantino (28,200,132). Then, I wanted to see what the average film rating was for each of these directors. Christopher Nolan's average film rating was 8.290244, Spielberg's was 7.066346, Lucas's was 7.079167, Cameron's was 7.721429, and Tarantino's was 7.789474 (**figure 2B**). All these director ratings fell close to the observed ratings from the histogram in figure 2A. When observing the boxplots for each director's film ratings the range of film ratings for each director did not appear abnormally large, but I found it interesting that George Lucas's films exhibited a large range of ratings. This is likely due to his Star Wars movies rating well whereas his earlier works, which were made before Lucas was well known, rated lower.

To understand why these directors were the most popular, I aimed to see what genres each of these directors was most associated with, as I hypothesized that genre may play a role in their popularity. However, this did not seem to be the case because there was a lot of variability in their top genres. Christopher Nolan has released mostly Drama films, Steven Spielberg mostly adventure, George Lucas released an equal number of action, adventure, and fantasy films, James Cameron mostly action, and Quentin Tarantino mostly crime films (**figure 2C**). Although certain genres did not seem to be associated with popularity, I wanted to see if a director's film rated lower when they released films outside of their preferred genre. For example, I observed if Christopher Nolan's non-drama films rated comparably to his drama films. Again, the genre did not seem to strongly relate to average rating. Nolan's drama films rated 8.476667, while films of other genres were comparable and slightly higher. Specifically, films of adventure, crime, or biography genres rated better with scores of 8.750000, 8.607143, and 8.600000 respectively. Overall, it seems that genre did not play a role into popularity, calling for further analysis of other factors that could help explain the popularity of these directors.

In conclusion, the United States, followed by Great Britain, India, Japan, and Canada released the most movies. Looking at US directors, Nolan, Spielberg, Lucas, and Cameron, received the most attention, or votes, in the IMDb database. Even though I thought genre might help explain their popularity, these top directors released films of many different genres, and director popularity and genre did not seem to be tightly associated. Nevertheless, this analysis illuminates top regions and directors to help the modern movie-lover, such as myself, to choose their next film to watch.

Figures

Figure 1A

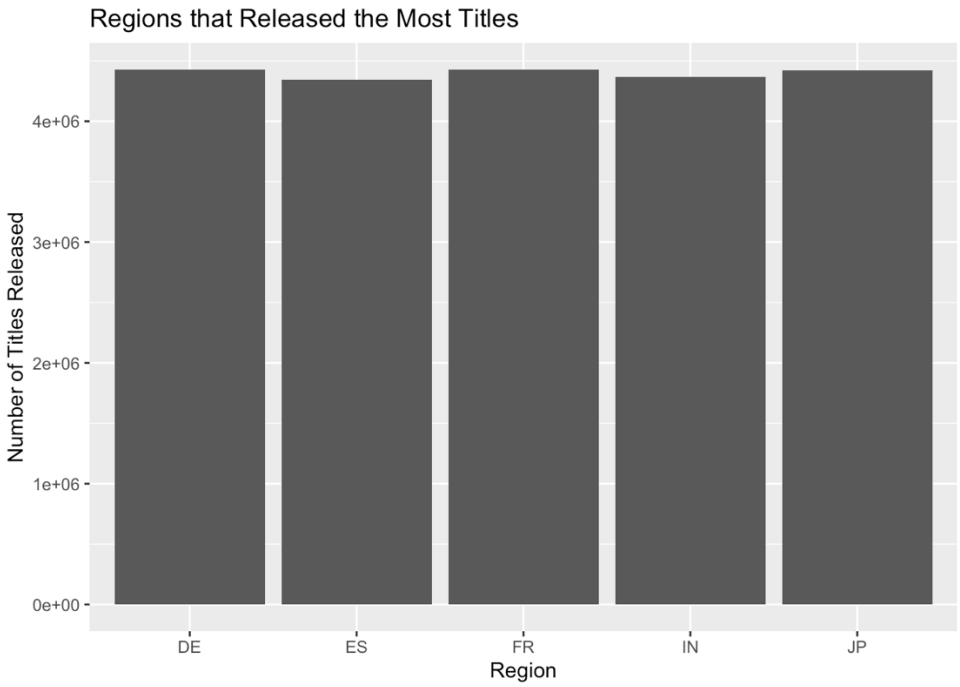


Figure 1B

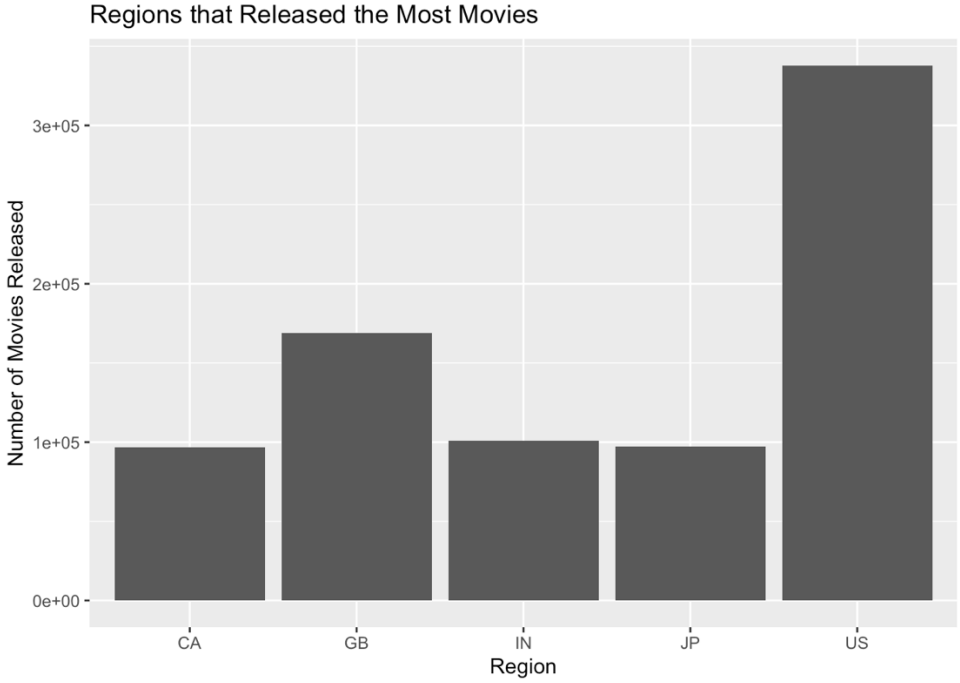


Figure 1C

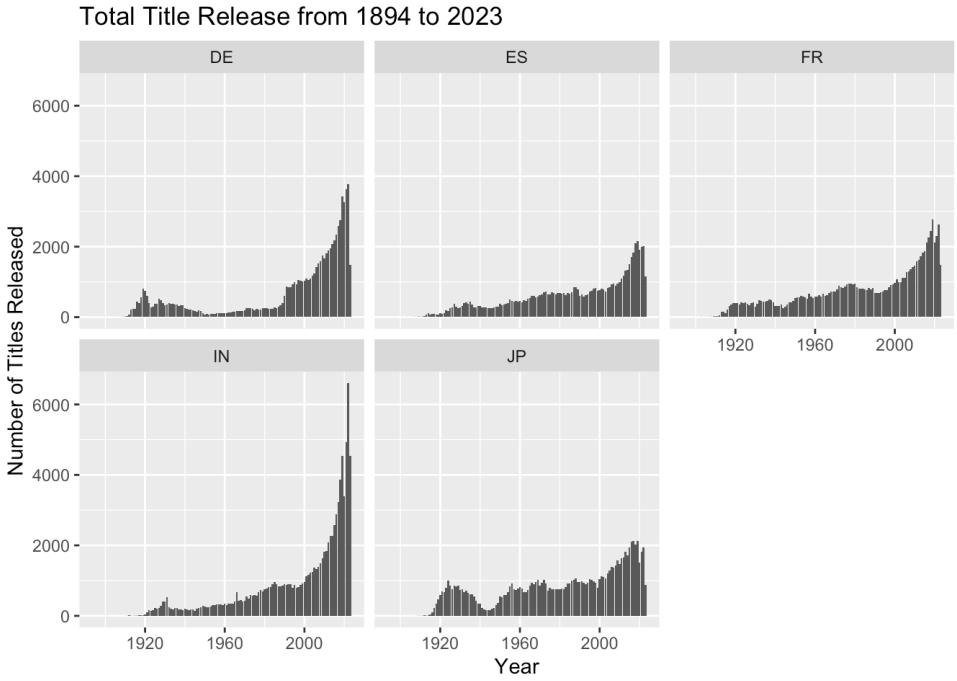


Figure 1D

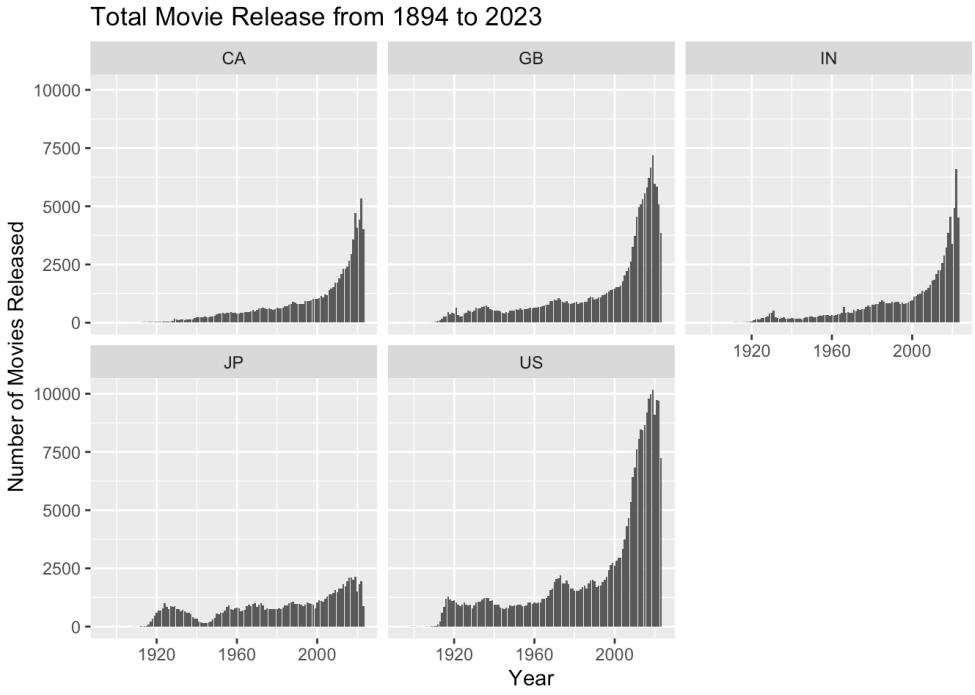


Figure 2A

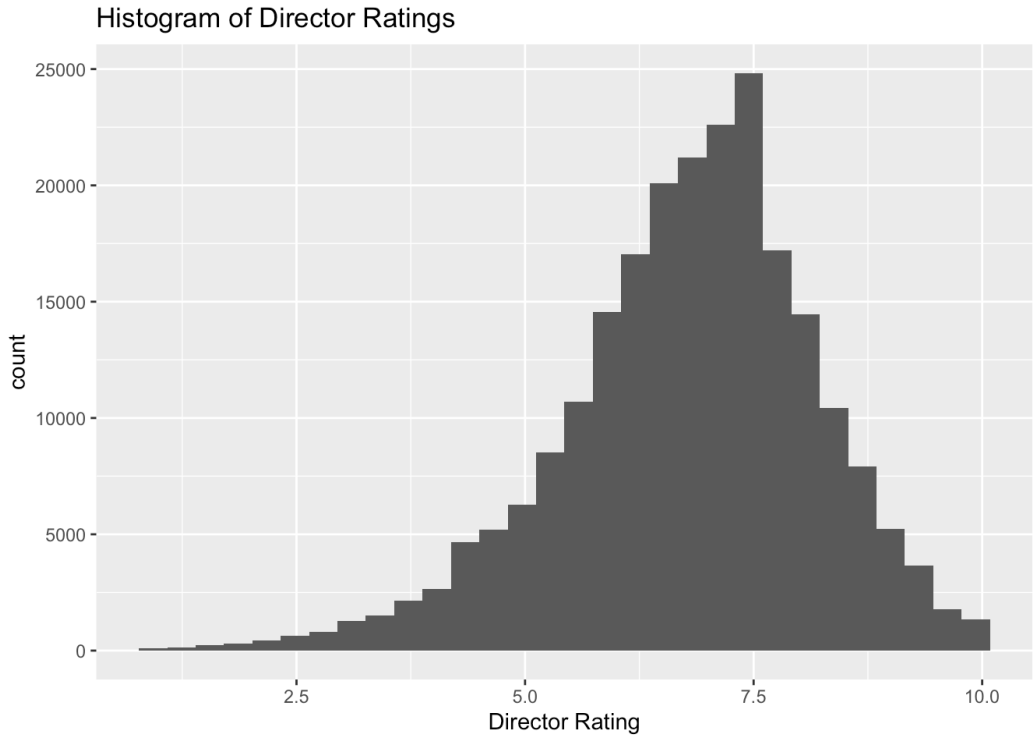


Figure 2B

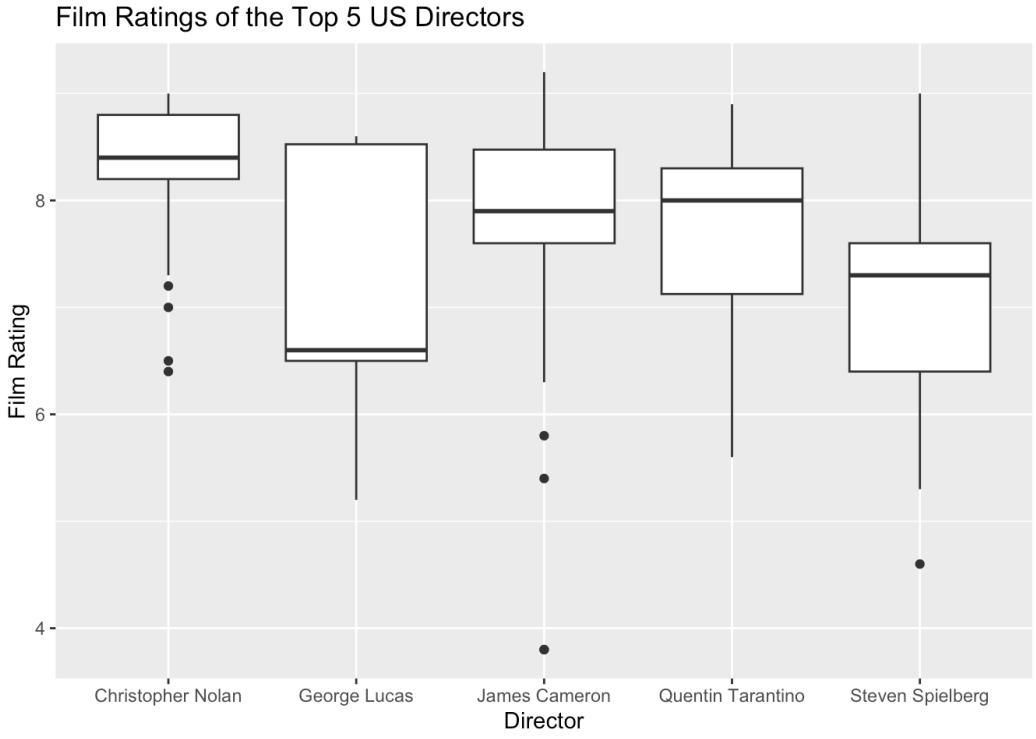
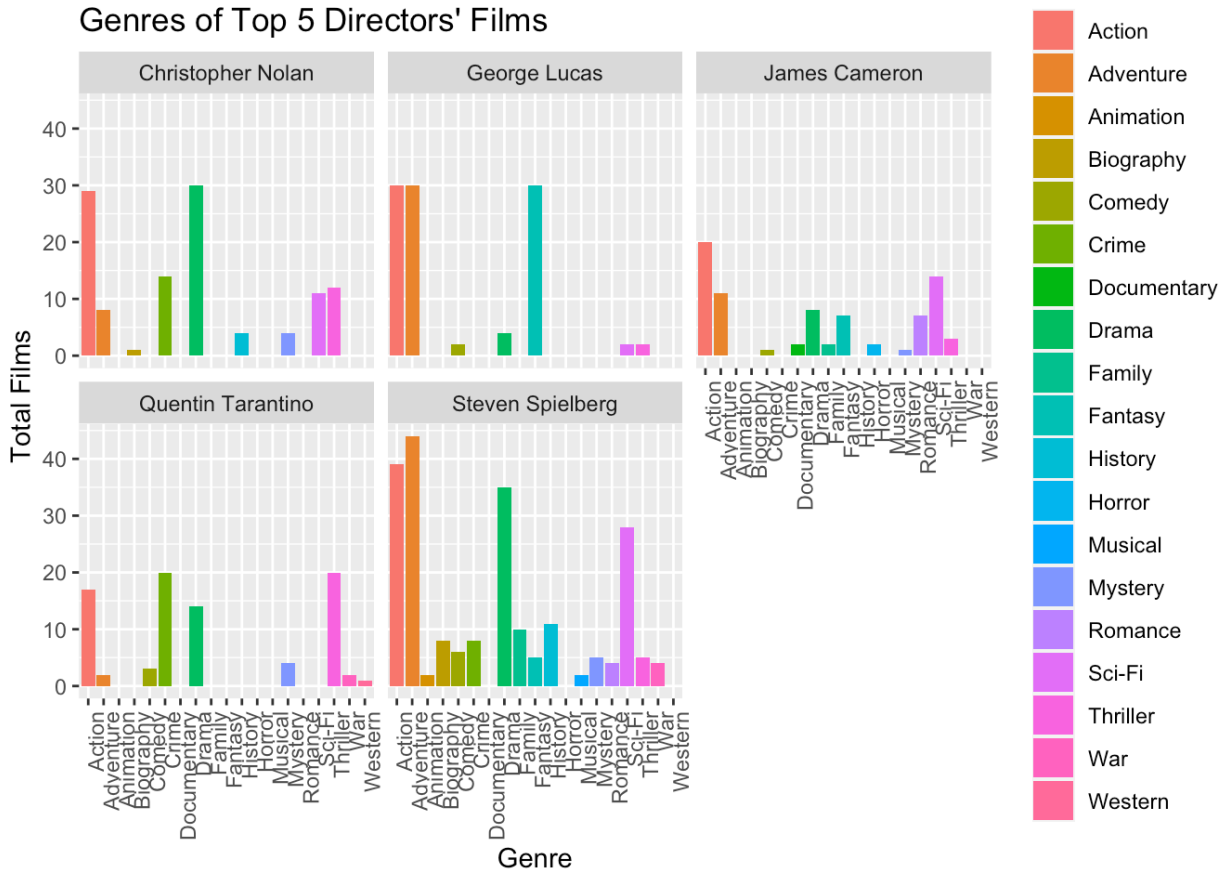


Figure 2C


Genres of Top 5 Directors' Films
Christopher Nolan
George Lucas
James Cameron
Quentin Tarantino
Steven Spielberg
Total Films
0
10
20
30
40
Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Musical
Mystery
Romance
Sci-Fi
Thriller
War
Western
Genre