Grace Techau

Data Wrangling Final Project Report - Fall 2024

Letterboxd Ratings and Box Office Revenue

1. Introduction

The popular social media platform Letterboxd was created in 2011 with a purpose of providing movie lovers an "appropriate venue to write about and share their love of film". As a long time user of Letterboxd, I have logged hundreds of movies with a rating out of 5 stars and personalized reviews. Letterboxd tracks all users ratings and calculates a weighted average rating for all movies in their database, as well as generic data for each film including the length and genre. For this project, I wanted to utilize the data on Letterboxd and conclude if a movie's external success at all contributes to the success of that movie on Letterboxd in the form of a higher rating. I determined the best way to measure a movie's base-line success is through its box office revenue - or the amount of money generated from ticket sales for a movie during its debut theatrical release.

By combining data on a movie's Letterboxd rating and box office revenue I want to answer the question: do movies that performed better in the box office also have a higher average rating on Letterboxd? I am also interested in seeing if any generic movie details impact average Letterboxd rating and/or box office revenue, including the length of a movie and the specific genre of a movie.

2. Data

This project uses two primary sources of data: a dataset titled 'Movies Box Office Collection Data 2000-2024' collected from collaborative data science website Kaggle and movie detail and rating information from the popular movie review website Letterboxd. I originally intended to collect data across both sources for movies released during the years 2011 to 2019. However, upon further research I discovered that the Letterboxd app was not released to the public until 2013 and took a few years to grow in popularity. According to Letterboxd's wikipedia page, "Letterboxd users collectively logged their 100 millionth film" in 2017. This information combined with the impracticality of scraping information on 157,090 films available on

Letterboxd for the years 2011 to 2019, makes using a time range of movies released from 2017 to 2019 the most appropriate for this project. This range will still provide an adequate sample size that will more accurately represent movie rating data on Letterboxd during the height of the application's popularity while still avoiding films released during and after the COVID-19 pandemic (2020-2024) which may have hindered their box office performance.

2.1 Box Office Revenue

To collect the relevant information for this project from the original Kaggel dataset, I created a separate notebook for cleaning box office revenue data. First, I downloaded the original CSV file from Kaggle with information on movies released from 2010 to 2024 (*2010-2024_Movies_Box_Ofice_Collection_raw.csv*). This original file included worldwide, domestic and foreign box office revenue for movies released during these years. After importing the CSV file as a data frame I made the following changes:

- Drop identification column 'Rank'
- Rename columns using snake case formatting
- Remove '%' and ',' characters from numeric columns and correct data types
- Convert domestic and foreign percentage columns to decimales on scale 0 to 1
- Change the scale of revenue columns to millions and round to 2 decimals
- Drop all rows from years 2010-2016 and 2020-2023
- Fix format issue where a dash character (-) in title column shows as a zero (0) using function *fix_dashes*

After making these changes, I saved the finalized clean box office revenue data frame to a CSV file (*box_office_revenue_2017.2019_clean.csv*) in the clean data folder.

2.2 Letterboxd

The Letterboxd website has a specific page where they list all the movies in their database by year and where certain viewing filters can be applied. For the scope of this project, I scraped the top 25% of movies sorted by popularity for the years 2017, 2018 and 2019 while applying the filter 'Hide Short Films'. The reason I scraped only the top 25% most popular movies was mainly due to the impracticality of scraping all movies for each year. The Letterboxd website has over 250 pages of movies for 2017, 2018 and 2019 while hiding short films. This is over 60,000 movies for each year and would take more time and processing power to scrape than this project requires. When just scraping the top 25% most popular movies while hiding short films, I still scraped over 10,000 films for all three years - which is more than sufficient for this project.

I scraped the movies for each year in an independent notebook (*00_letterboxd.YEAR.scrape.ipynb*), but used the same methods and script. I used the Selenium package and a variety of user-defined functions in Python to scrape individual film links from the films by year page on Letterboxd and then scrape variables from each movie detail page.

The *apply_filters* function was utilized to access the filters drop down menu on the first Letterboxd films' page being scraped and click 'Hide short films'. This filter continued to apply to all pages scraped thereafter, and is why the function is only applied to the first page in the range being scraped. The *scrape_movie_links* function adds the URL link to each individual movie detail page for every movie on the page to a defined list. Finally, the *scrape_movie_pages* is a specific function that uses the *apply_filters*, *scrape_movie_links*, and *random_scroll* functions to crawl the 25% most popular pages of Letterboxd films for a certain year.

After scraping the individual movie detail page URLs from the main Letterboxd page while applying the filter 'Hide short films'. I then scraped through all movies and extracted the variables *title, year, average_rating, number_ratings, length,* and *genres* using CSS Selector with Selenium. For each year, the movie information was scraped in up to four batches which were all saved to a separate CSV file in the raw_data folder (*letterboxd_movie_data_YEAR_raw_X.csv*).

After scraping movie details for the 25% most popular movies on Letterboxd excluding short films, each raw data set was cleaned for the years 2017, 2018, and 2019. First, I merged the raw data files from scraping into one data frame for cleaning. Movies for each year were cleaned in separate notebooks (*02 - 04_letterboxd.YEAR.clean.ipynb*), but used the same methods and script. The following changes were made to the raw Letterboxd data while cleaning:

- Duplicate rows were dropped (existed due to the batch scraping)
- Rows with no average rating and no number of ratings available were dropped (irrelevant for analysis)

- Strip *length* column to remove text 'mins More at IMDB TMDB' and keep just numerical value of move length in minutes
- Strip *number_ratings* column to remove text and only include numerical value of total number of ratings without comma
- Remove any duplicates of genres for same movie and only keep first 3 genres for each movie
- Fix all data types

After making the following changes, I saved the finalized Letterboxd movie details data frame to a CSV file (*letterboxd_movie_data_YEAR_clean.csv*) in the clean_data folder for each year.

2.3 Combining Letterboxd and Box Office Revenue

After gathering and cleaning both the box office revenue and Letterboxd movie data, it was time to merge all files and make the finalized dataset for this project. First, I merged the clean CSV files for Letterboxd data across the years 2017, 2018 and 2019. I dropped all duplicates from this final Letterboxd datafile and then used an inner join to merge the clean box office revenue data on the features *title* and *year*. After merging, the final data frame has 11 columns and 491 rows. The features of the final box office revenue and Letterboxd rating dataset for the years 2017 to 2019 can be seen in the data dictionary below.

Table 1 - Data Dictionary

Column	Type	Source	Description
title	Text	Both	Title of the movie
year	Numeric	Both	Year the movie was released
worldwide_revenue	Numeric	Kaggle	Revenue in U.S. dollars that the movie made at the box office worldwide (in millions)
domestic_revenue	Numeric	Kaggle	Revenue in U.S. dollars that the movie made at the box office in the United States and Canada (in millions)
domestic_percent	Numeric	Kaggle	Percentage of worldwide box office revenue that is attributed to domestic box office revenue

foreign_revenue	Numeric	Kaggle	Revenue in U.S. dollars that the movie made at the box office in foreign countries (in millions)
foreign_percent	Numeric	Kaggle	Percentage of the worldwide box office revenue that is attribute to foreign box office revenue
number_ratings	Numeric	Letterboxd	Total count of ratings the movie has by Letterboxd users
average_rating	Numeric	Letterboxd	Average star rating of the movie out of 5
length	Numeric	Letterboxd	Length of the movie in minutes
genre	Text	Letterboxd	List of up to the 3 most relevant genres of the movie

3. Analysis

3.1 EDA

To understand this new data more thoroughly, I developed a variety of exploratory data analysis visualizations. These are shown in Figures 1 through 12 below.

Figure 1 - Average Rating of ~3 is Most Common on Letterboxd

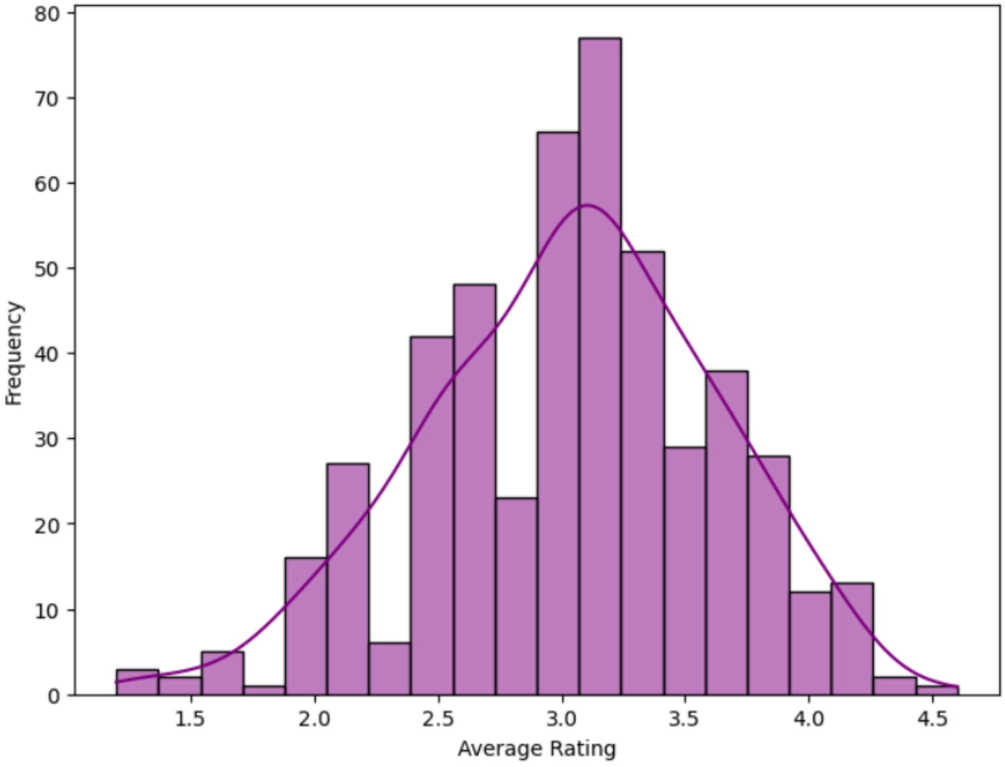


Figure 2 - Right Skew Distribution of Worldwide Revenue

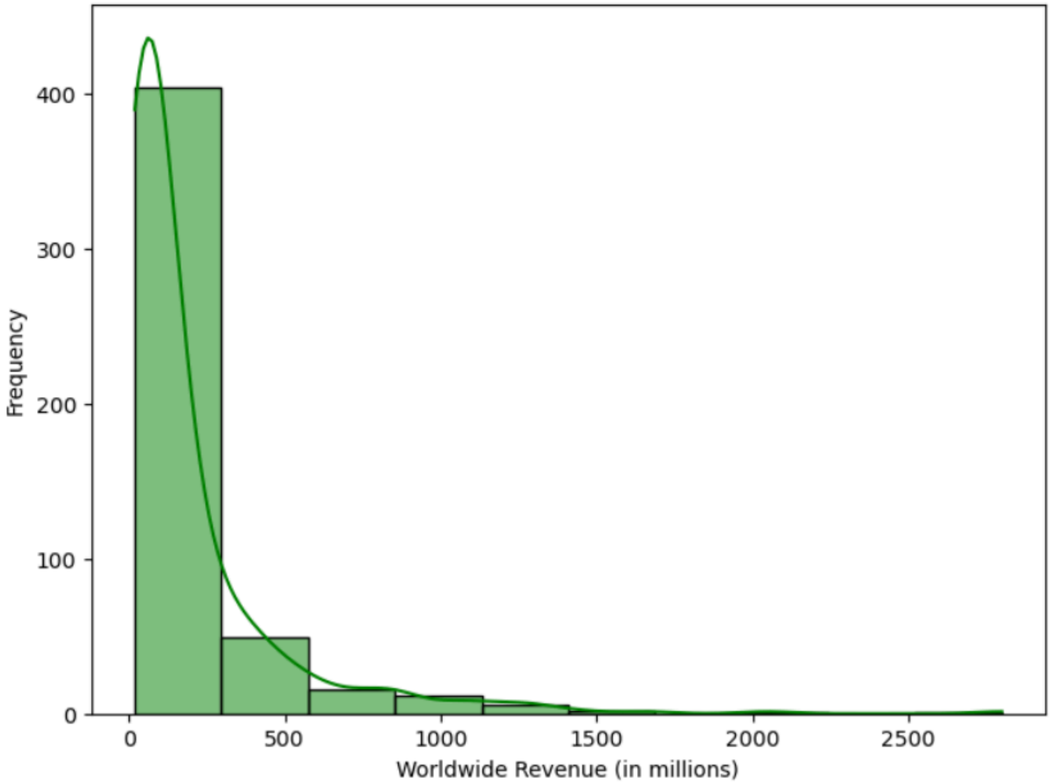


Figure 3 - Slight Right Skew Distribution of Film Length

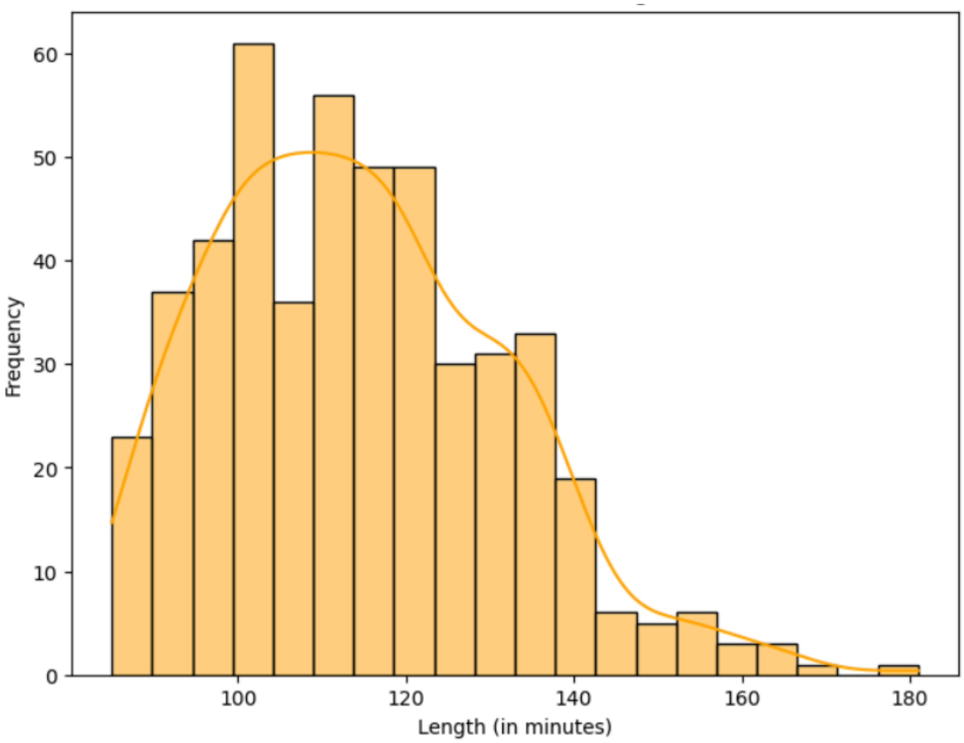


Figure 4 - Correlation Matrix of Numerical Features

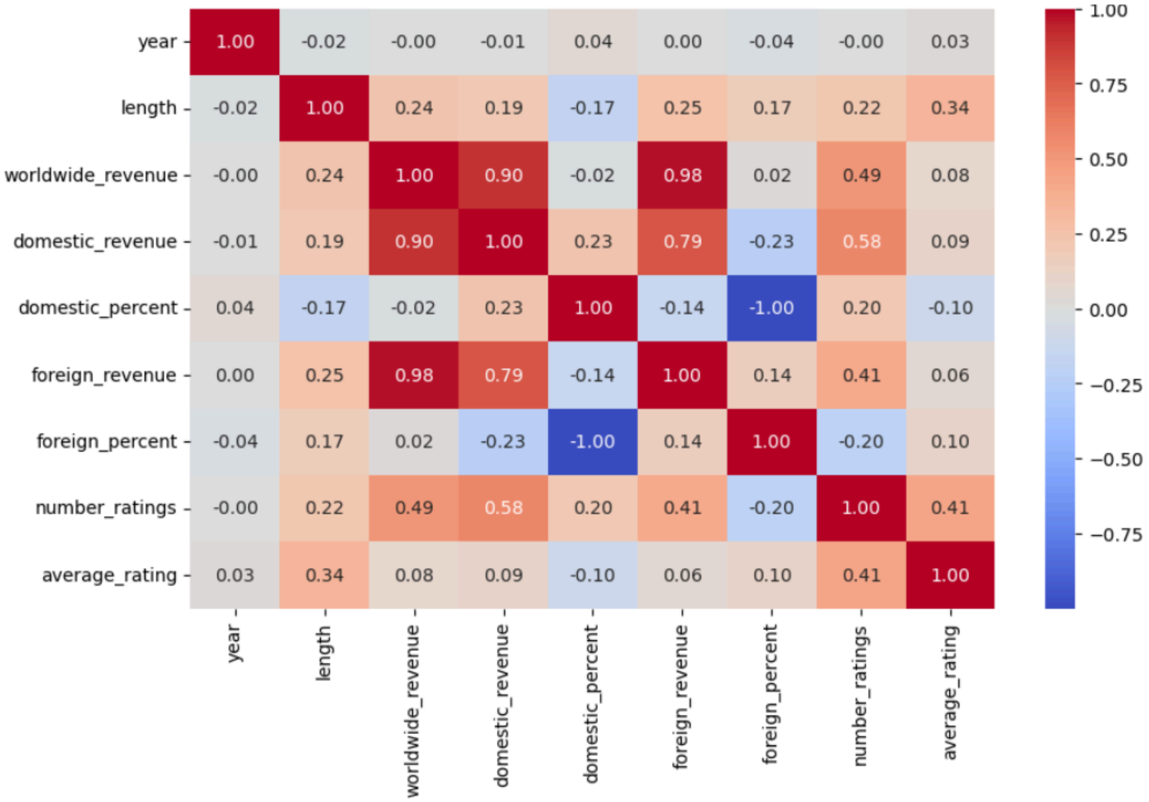


Figure 5 - Movie Length v. Average Rating

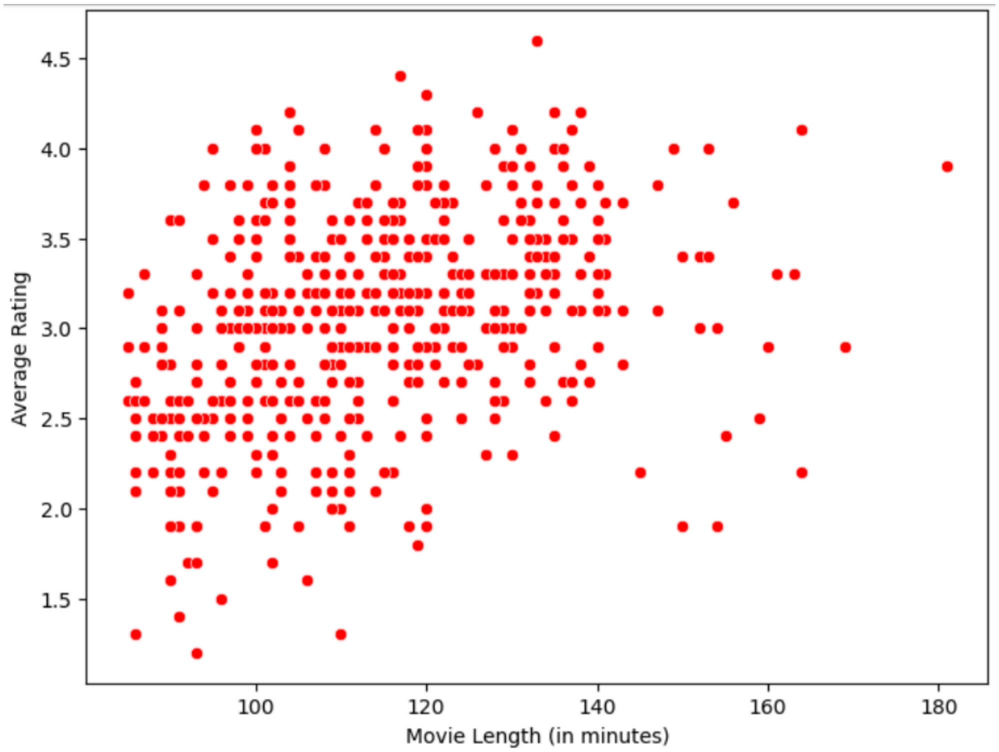


Figure 6 - Movie Length v. Worldwide Revenue

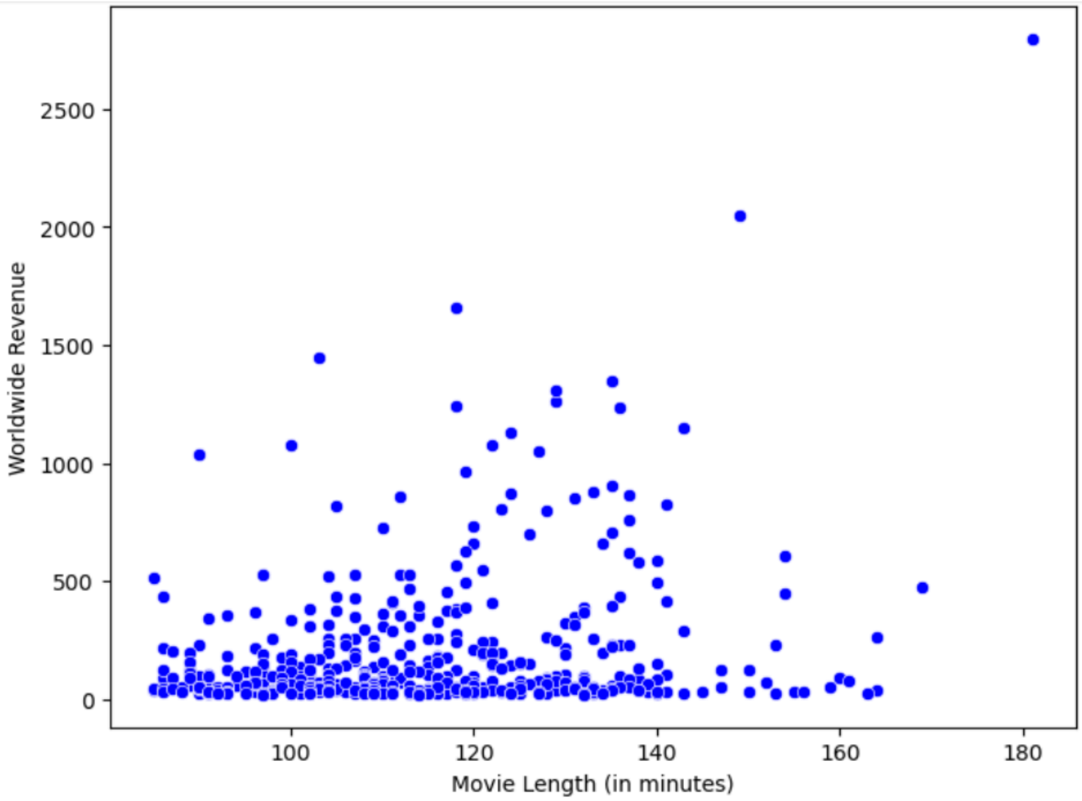


Figure 7 - Domestic Revenue v. Foreign Revenue

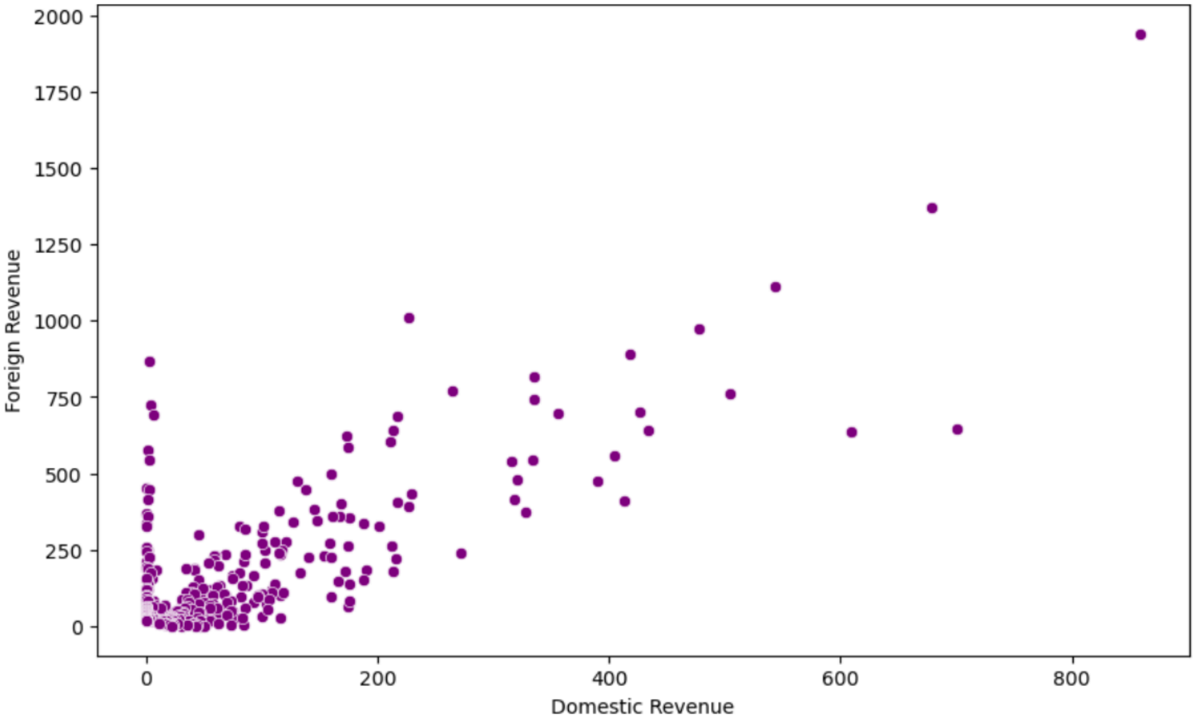


Figure 8 - Top 10 Most Common Genres

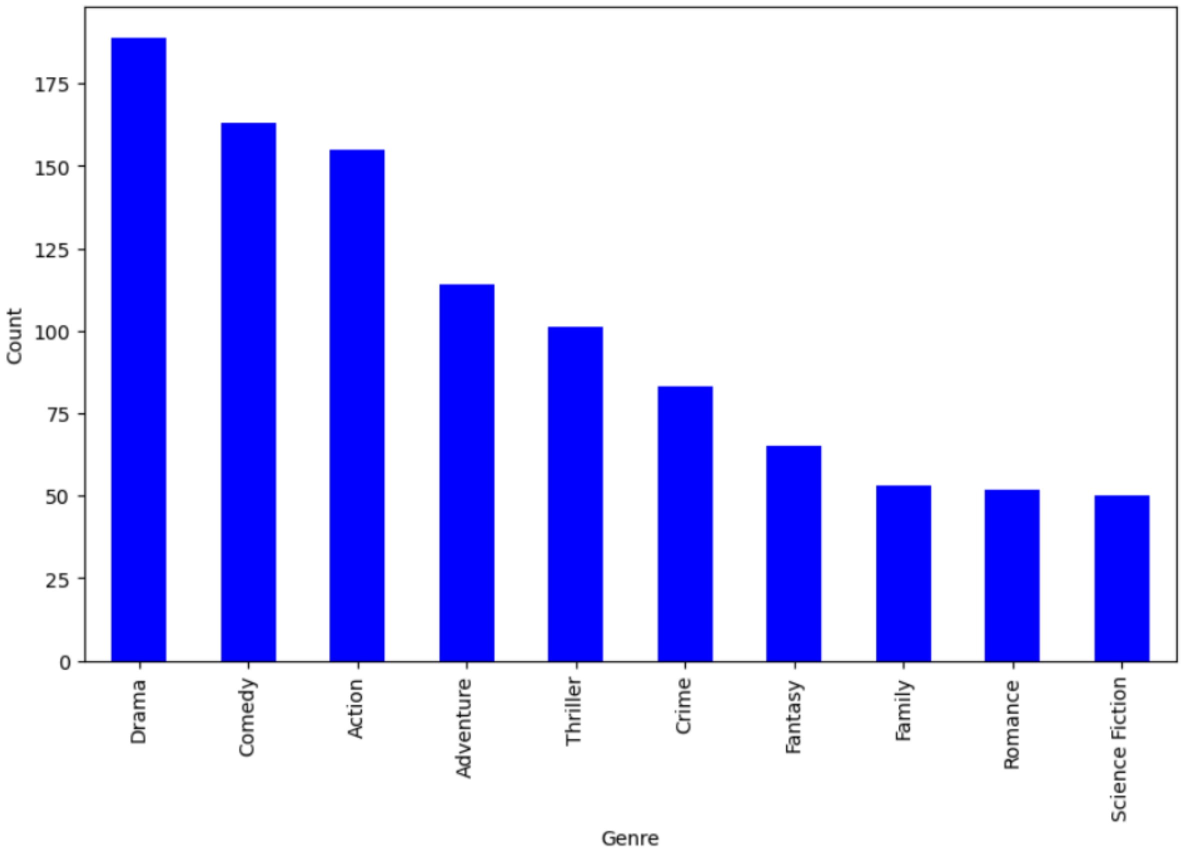
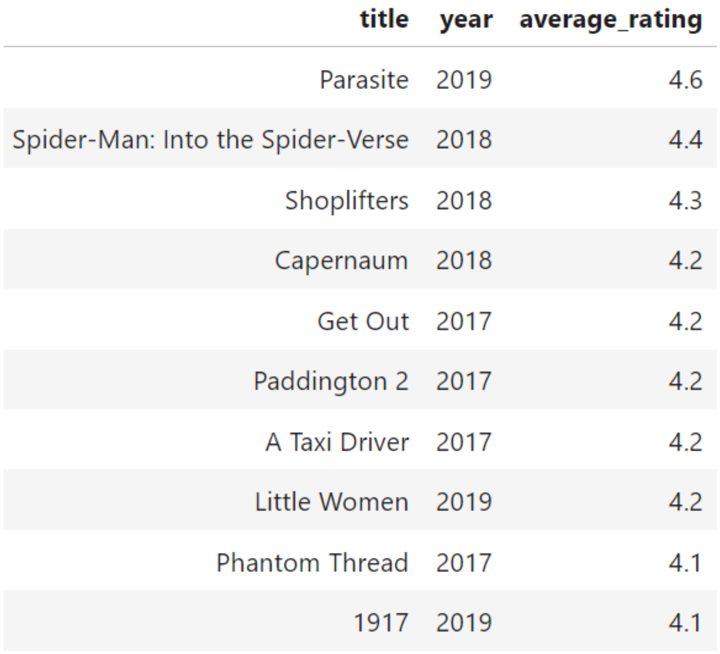
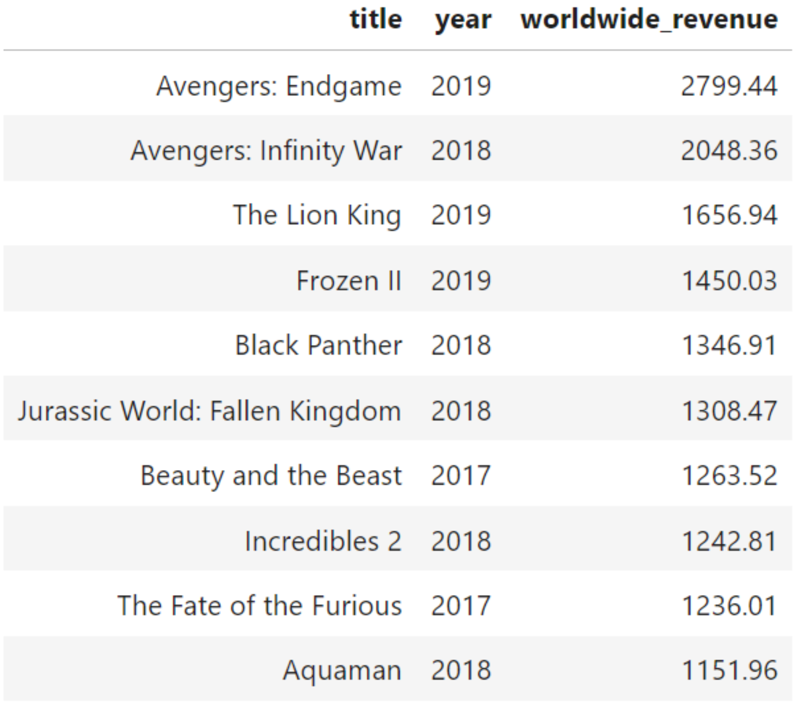


Figure 9 - Parasite is Highest Rated Movie on Letterboxd Years 2017-2019



title	year	average_rating
Parasite	2019	4.6
Spider-Man: Into the Spider-Verse	2018	4.4
Shoplifters	2018	4.3
Capernaum	2018	4.2
Get Out	2017	4.2
Paddington 2	2017	4.2
A Taxi Driver	2017	4.2
Little Women	2019	4.2
Phantom Thread	2017	4.1
1917	2019	4.1

Figure 10 - Avengers Series are Highest Earning Movies based on Worldwide Box Office Revenue Years 2017-2019



title	year	worldwide_revenue
Avengers: Endgame	2019	2799.44
Avengers: Infinity War	2018	2048.36
The Lion King	2019	1656.94
Frozen II	2019	1450.03
Black Panther	2018	1346.91
Jurassic World: Fallen Kingdom	2018	1308.47
Beauty and the Beast	2017	1263.52
Incredibles 2	2018	1242.81
The Fate of the Furious	2017	1236.01
Aquaman	2018	1151.96

Figure 11 - Average Letterboxd Rating has Grown Steadily 2017 to 2019

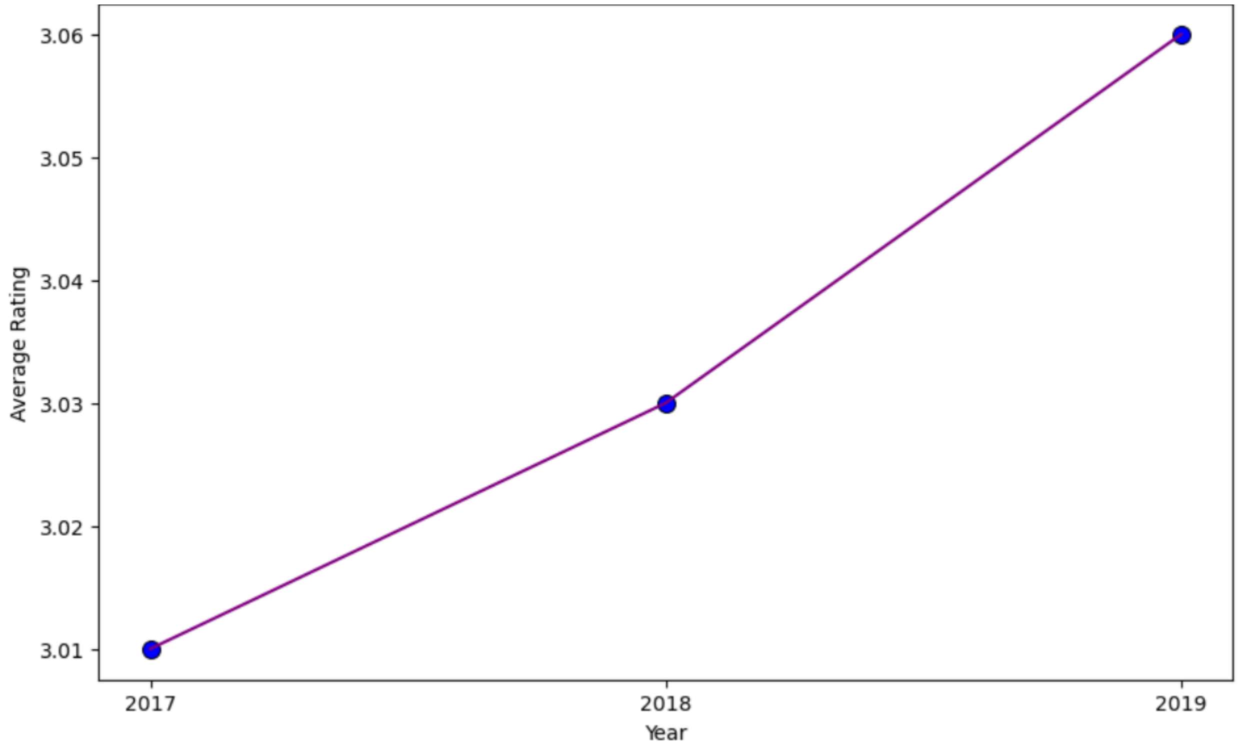
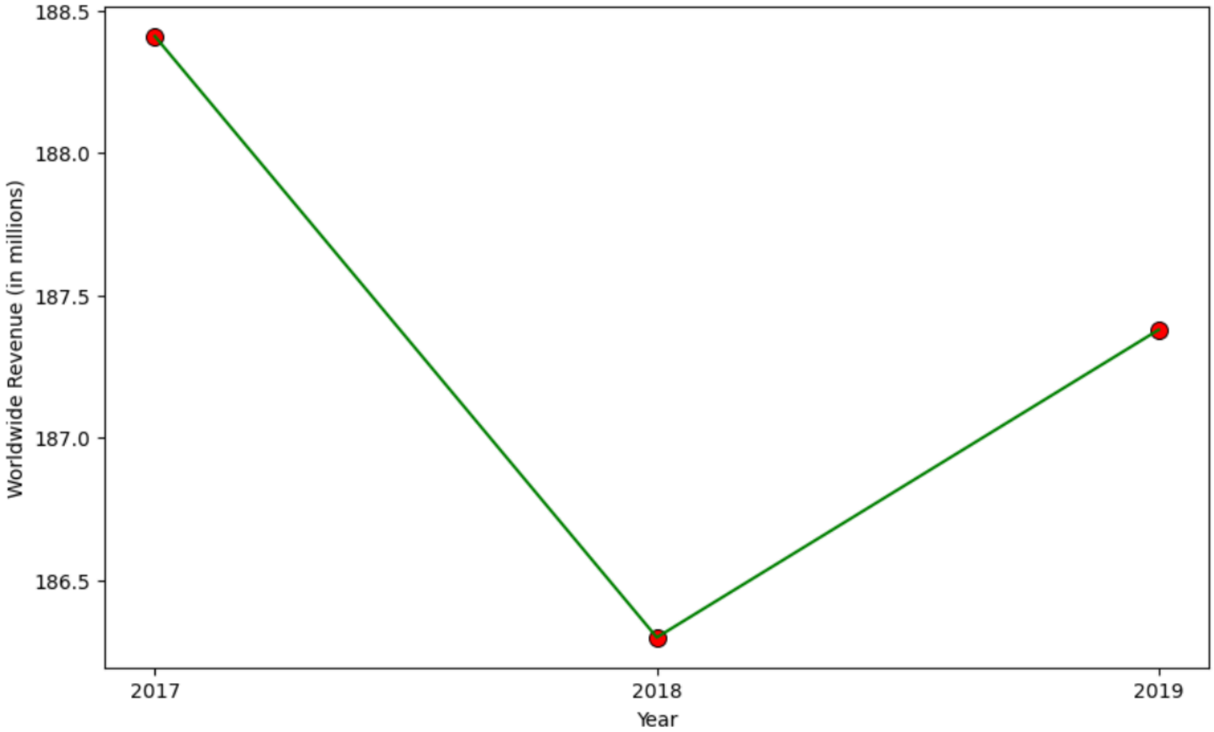


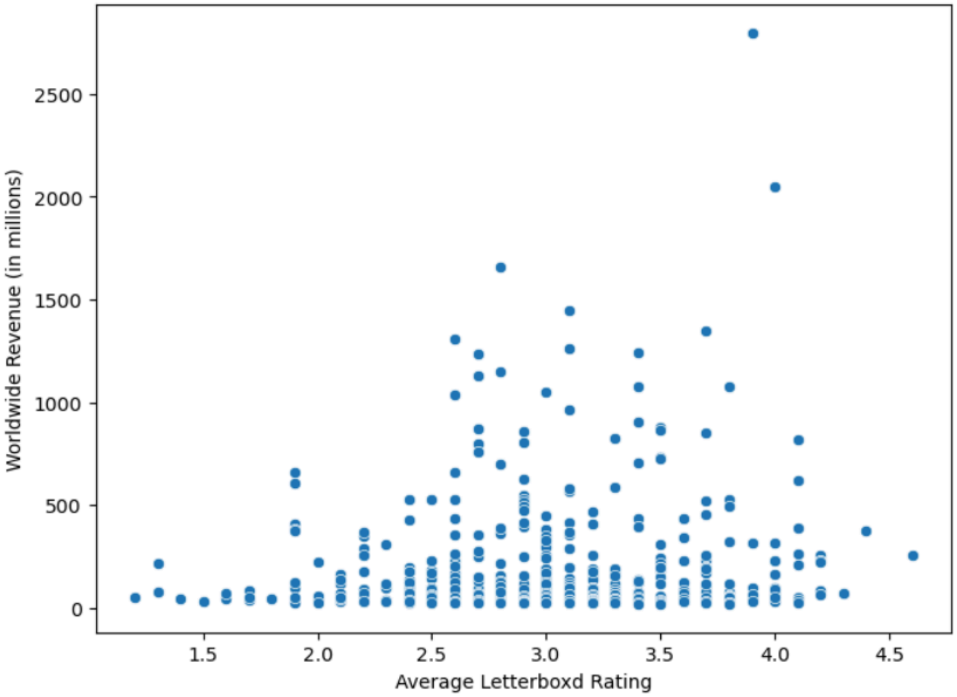
Figure 12 - Average Worldwide Revenue (in millions) Dropped in 2018



3.2 Box Office Revenue and Average Rating

A major theme throughout my project, and one of my main analysis questions, was to determine whether or not there is a correlation between worldwide box office revenue and average rating on Letterboxd. To determine if this is true or not, I first completed a Pearson correlation between *worldwide_revenue* and *average_rating* using scipy.stats from the pearsonr Python package. The Pearson correlation was 0.0763 which is less than 0.3 meaning the interpretation is ‘No or weak linear correlation’. This is further exemplified in Figure 13 which displays a weak relationship between a movie’s average rating on Letterboxd and its worldwide box office revenue. There is a very slight positive linear correlation, indicating that movies that made more worldwide box office revenue may have a slightly higher Letterboxd rating. However, the extremely weak Pearson correlation makes it most accurate to conclude that average rating on Letterboxd and worldwide box office revenue are not directly correlated.

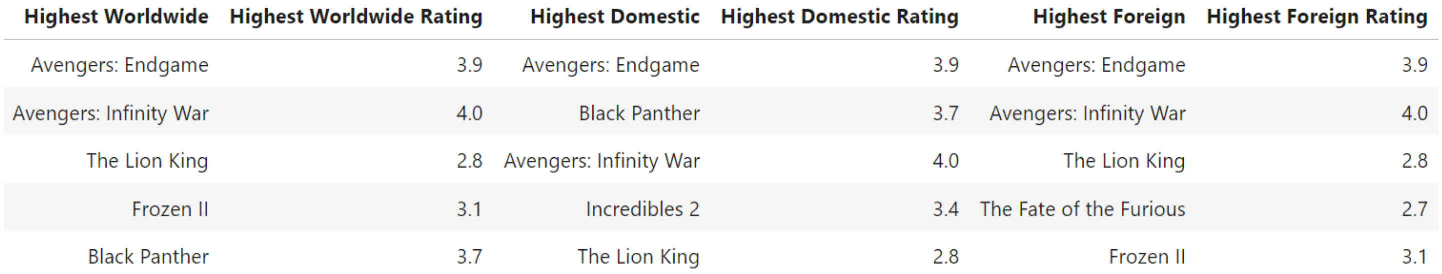
Figure 13 - Worldwide Revenue v. Average Rating



I also wanted to more specifically examine if movies with higher domestic and/or foreign revenue have higher average Letterboxd rating. I performed a Pearson correlation between the variables *foreign_revenue* and *domestic_revenue* to *average_rating*. Similarly to worldwide revenue, both domestic and foreign revenue have a weak relationship to average rating on Letterboxd and had a final interpretation of 'No or weak linear correlation'. The Pearson correlation between *domestic_revenue* and *average_rating* was 0.0928 and *foreign_revenue* and *average_rating* was 0.0623.

To finalize the conclusion that there is no significant relationship between box office revenue and average rating on Letterboxd, I also queried the 5 highest earning movies in terms of worldwide, domestic, and foreign revenue and their associated Letterboxd rating. The results are shown in Figure 14 and while all the highest earning movies had an average rating of above 2.5, there is no evidence that they are particularly higher than any other films in the data set.

Figure 14 - Films with Highest Revenue and Associated Letterboxd Rating



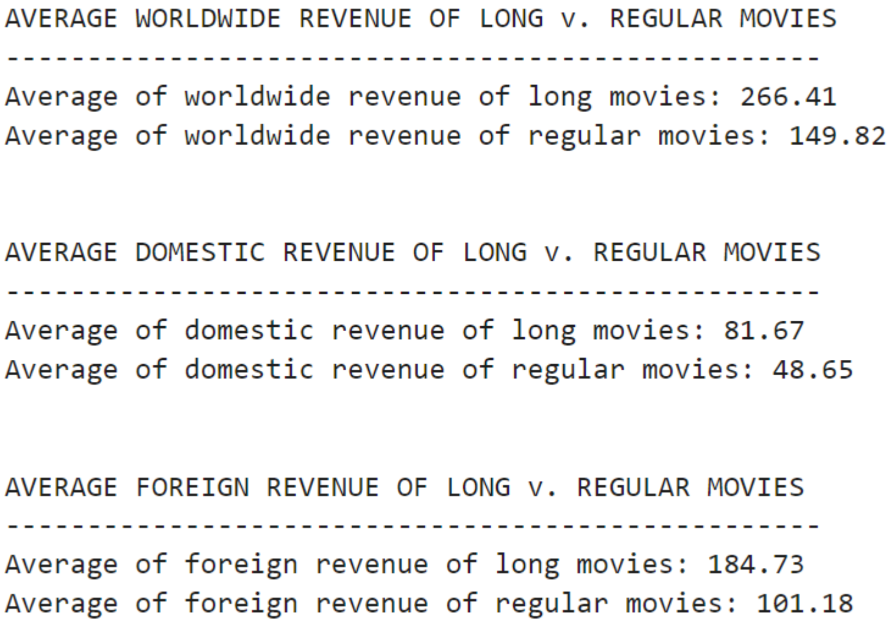
Highest Worldwide	Highest Worldwide Rating	Highest Domestic	Highest Domestic Rating	Highest Foreign	Highest Foreign Rating
Avengers: Endgame	3.9	Avengers: Endgame	3.9	Avengers: Endgame	3.9
Avengers: Infinity War	4.0	Black Panther	3.7	Avengers: Infinity War	4.0
The Lion King	2.8	Avengers: Infinity War	4.0	The Lion King	2.8
Frozen II	3.1	Incredibles 2	3.4	The Fate of the Furious	2.7
Black Panther	3.7	The Lion King	2.8	Frozen II	3.1

3.3 Influence of Movie Length

Some of the questions I proposed had to do with determining if the length of a movie impacts its box office revenue and/or Letterboxd average rating.

I found the average worldwide, domestic, and foreign revenue for both “long” and “regular” movies. I queried “long” movies as movies that were more than 120 minutes (2 hours) in length. The results I found in Figure 15 shows that long movies typically had a higher average worldwide, domestic and foreign revenue. Long movies had a higher worldwide revenue by 116.59 million, higher domestic revenue by 33.02 million, and higher foreign revenue by 83.55 million over regular length movies (less than 2 hours).

Figure 15 - Box Office Revenue of Long v. Regular Movies



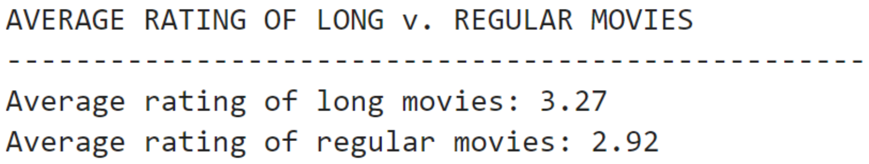
```
AVERAGE WORLDWIDE REVENUE OF LONG v. REGULAR MOVIES
--------------------------------------------------
Average of worldwide revenue of long movies: 266.41
Average of worldwide revenue of regular movies: 149.82


AVERAGE DOMESTIC REVENUE OF LONG v. REGULAR MOVIES
--------------------------------------------------
Average of domestic revenue of long movies: 81.67
Average of domestic revenue of regular movies: 48.65


AVERAGE FOREIGN REVENUE OF LONG v. REGULAR MOVIES
--------------------------------------------------
Average of foreign revenue of long movies: 184.73
Average of foreign revenue of regular movies: 101.18
```

I also determined the average Letterboxd rating for long and regular movies which had similar results as box office revenue. As seen in Figure 16, long movies (over 2 hours) have a higher average Letterboxd rating compared to regular movies.

Figure 16 - Average Rating of Long v. Regular Movies

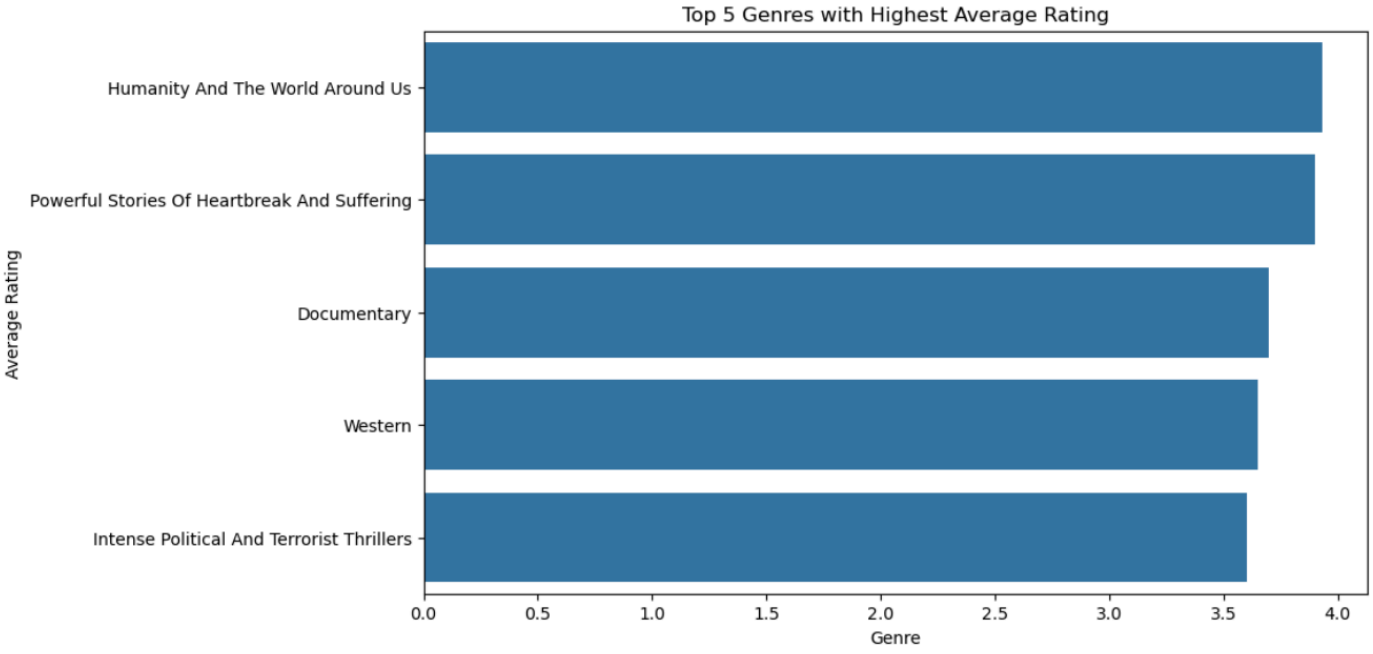


```
AVERAGE RATING OF LONG v. REGULAR MOVIES
--------------------------------------------------
Average rating of long movies: 3.27
Average rating of regular movies: 2.92
```

3.4 Genre Information

To gather information on genre data, I had to iterate through all rows in the film data and split the lists of genres for each movie by a comma delimiter. Then, I looped through each list of genres for each movie and gathered the average rating, genre name, and count of each distinct genre. Finally, I calculated the average rating for each of the 49 distinct genres across all movies based on the number of ratings for each genre. The five genres with the highest average Letterboxd rating can be seen in Figure 17.

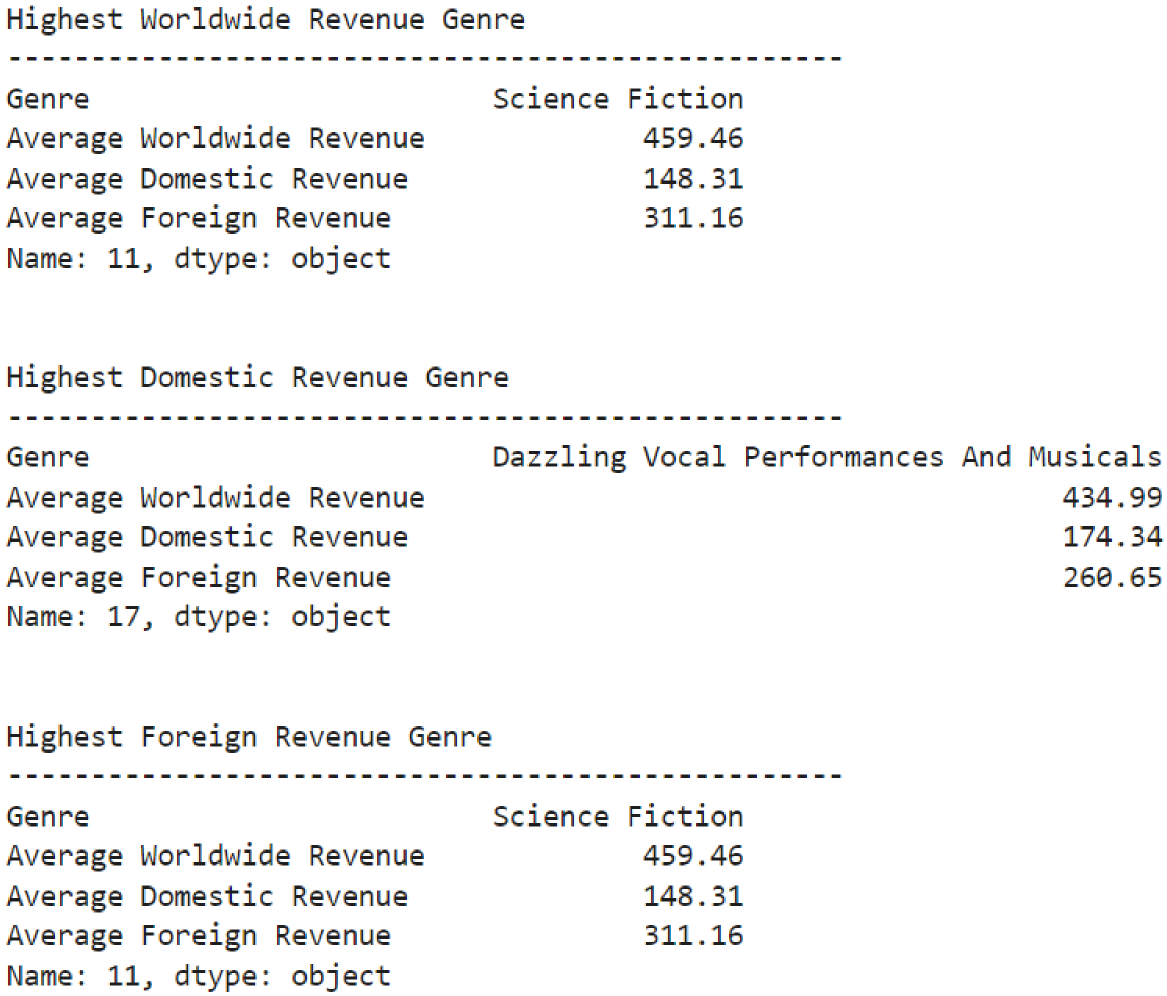
Figure 5 - Top 17 Genres with Highest Average Rating



The genre with the highest average rating is Humanity and The World Around Us with an average Letterboxd rating of 3.93 and the genre with the lowest average rating is Thrillers and Murder Mysteries with an average Letterboxd rating of 1.8.

I also gathered information relating to average box office revenue for the different genres. The genre with the highest average worldwide and foreign box office revenue is Science Fiction with 459.46 million and 311.16 respectively. The genre with the highest average domestic box office revenue is Dazzling Vocal Performance and Musicals with 174.34 million. The average box office revenue data for all of the highest earning genres can be seen in Figure 18.

Figure 18 - Genres with Highest Worldwide, Domestic and Foreign Revenues



```
Highest Worldwide Revenue Genre
--------------------------------------------------
Genre                            Science Fiction
Average Worldwide Revenue                 459.46
Average Domestic Revenue                  148.31
Average Foreign Revenue                   311.16
Name: 11, dtype: object


Highest Domestic Revenue Genre
--------------------------------------------------
Genre                            Dazzling Vocal Performances And Musicals
Average Worldwide Revenue                                          434.99
Average Domestic Revenue                                           174.34
Average Foreign Revenue                                            260.65
Name: 17, dtype: object


Highest Foreign Revenue Genre
--------------------------------------------------
Genre                            Science Fiction
Average Worldwide Revenue                 459.46
Average Domestic Revenue                  148.31
Average Foreign Revenue                   311.16
Name: 11, dtype: object
```

Conclusion

In this project, I focused on the possible correlation between a film's box office revenue and average Letterboxd rating. More specifically, I searched for the answers to the analysis questions presented in my proposal and found the following results.

1. *Do movies that performed better at the box office in terms of worldwide, domestic and foreign revenue, have a higher average rating on Letterboxd?*

   Based on the analysis in section 3.2 and the correlation coefficient of 0.0763, it can be concluded that there is no correlation between box office revenue, in terms of worldwide, foreign or domestic, to average rating on Letterboxd.

2. *Do movies that have higher foreign or domestic revenue have a higher average rating on Letterboxd?*

   As the work in section 3.2 shows, the correlation between foreign and domestic revenue to average Letterboxd rating are both extremely weak. However, domestic revenue does have a slightly higher correlation to average rating (0.0928) compared to foreign revenue (0.0623). This means that movies with higher domestic box office revenue are slightly more likely to have a high average rating on Letterboxd compared to a movie with high foreign box office revenue.

3. *Do longer or shorter movies make a higher revenue (worldwide, domestic and foreign) at the box office?*

   Based on analysis done in section 3.3, longer movies - or movies longer than 120 minutes in length - have a higher average worldwide, domestic and foreign revenue compared to regular length movies.

4. *Do longer or shorter movies get rated higher on Letterboxd?*

   Similarly to question 3 and analysis done in section 3.3, longer movies also had a higher average rating on Letterboxd compared to regular length movies, or movies less than 120 minutes in length.

5. *What genre of movie has the highest ratings on Letterboxd?*

   The genre with the highest average rating on Letterboxd is Humanity And The World Around Us with an average rating of 3.93 followed by Powerful Stories of Heartbreak And Suffering with an average rating of 3.90. These findings are exemplified further in section 3.4.

6. *Does that same genre of movie perform better at the box office in terms of worldwide, domestic and foreign revenue?*

   Based on analysis done in section 3.4, the genre with the highest average Letterboxd rating (Humanity And The World Around Us) does *not* have the highest worldwide, domestic or foreign revenue. Instead, the genre Science Fiction had the highest average worldwide and foreign box office revenue. The genre Dazzling Vocal Performances and Musicals had the highest average domestic box office revenue.

The major limitation of this project was the time constraint to scrape movies from the Letterboxd website. I believe that the information gathered from both sources was accurate and useful for drawing conclusions about a film's box office revenue and Letterboxd rating information. However, only having data on movies from 2017 to 2019 could have limited potential correlations between box office revenue and Letterboxd ratings over a longer period of time. As I explained in section 2, it was impractical for this project to scrape more movies. For future work I could scrape all movies on Letterboxd from 2013 to 2019 and potentially scrape, instead of using a pre-made dataset, box office revenue data for more movies during these years. This work could provide data on more movies in the final dataset and therefore have more accurate analysis and meaningful conclusions.