Analysis and forecast of the film industry

Yuntian Shi

Summary of research questions:

1. How does a movie's budget relate to its box office success? Are there specific budget ranges that tend to result in the highest return on investment?

2. How does the reputation or track record of a director impact a movie's success? Are there specific directors whose involvement in a project tends to increase the likelihood of critical or commercial success? Does a director's past performance or success influence the types of movies they are able to make or the budgets they are given?

3. Do movie reviews differ depending on the gender of the audience? And are there significant differences in the kinds of movies men and women enjoy?

4. If I want to train a regression model to predict future movie revenue, which variables in a movie industry dataset have the greatest predictive power for a movie's future revenue? Can a model be developed that accurately predicts revenue based on factors such as a movie's genre, production budget, or release date?

Motivation:

As a college student with a passion for film and a budding interest in data analysis, my motivation for investigating the movie industry dataset is multi-faceted. Firstly, I am interested in gaining a deeper understanding of the movie industry as a whole, and exploring the various factors that contribute to a film's success, both commercially and critically. By delving into this dataset, I hope to uncover patterns and insights that can help me better understand the movie-making process and the challenges and opportunities that exist within the industry. Additionally, as a student of data analysis, I am eager to apply my skills to a real-world dataset and gain hands-on experience with techniques such as data cleaning, visualization, and predictive modeling. Finally, I believe that exploring the movie industry dataset will be an enjoyable and engaging way to apply my interests and skills to a project that has the potential to yield real-world insights and impact.

Dataset:

The source of my dataset is Kaggle. So far I found three datasets, the first one has 6820 movies (220 movies per year, 1986-2016) with 15 different attributes. The second dataset contains metadata for about 5,000 movies from TMDb. The third dataset is the top 1000 films rated by IMDB. These datasets all have a common attribute (movie name).

- https://www.kaggle.com/datasets/danielgrijalvas/movies
- https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
- https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Challenge goals:

- **Multiple Datasets:** I intend to use three or more datasets in my research problem, by combining the datasets in a way that I can get more information to use for analysis.
- **Machine Learning:** I want to train a model that can predict movie revenue. This may be a bit difficult, I just have a general idea at the moment, not sure of the exact steps yet.

Method:

Q1:
Combine the three movie industry datasets and select the relevant variables that are related to a movie's budget and box office performance.
Check the datasets for missing or incomplete data, and clean the data as necessary to ensure that the data is accurate and complete.
Create a variable that represents the ratio of a movie's budget to its box office gross.
Calculate the average value of this variable and randomly select a movie for testing.
Check the error range.

Q2:
Groupby directors and calculate their potential by analyzing the number of films and the percentage of box office generated by each film.

If a director has multiple productions in 20 years, 10 years is used as a threshold to predict the success of his work in the next 10 years compared to the success of his work 10 years ago.
The success of a film is judged by its rating, the number of votes and the gross.
Check if the budget, rating, and box office performance of those directors who did well 10 years ago are as good as expected 10 years later.

Q3:
The budget is used to determine if an actor is a famous star.
Filter out the data of those who are not big names, and female directors (use gender attribute).
Use the ratio of box office numbers to budget for the remaining data to analyze and determine whether the film has been a success in the marketplace.

Q4:
Use the LinearRegression model.
Features are the variables defined in the first problem.
Labels are gross.

Each question will be given a visualization to help explain it.

Work Plan:

Step 1: Data Preparation and Exploratory Analysis (6 hours)
Clean the dataset, removing any irrelevant or missing data.
Perform exploratory analysis to gain a better understanding of the variables and relationships within the dataset.
Create classes and methods that attempt to answer the research questions.

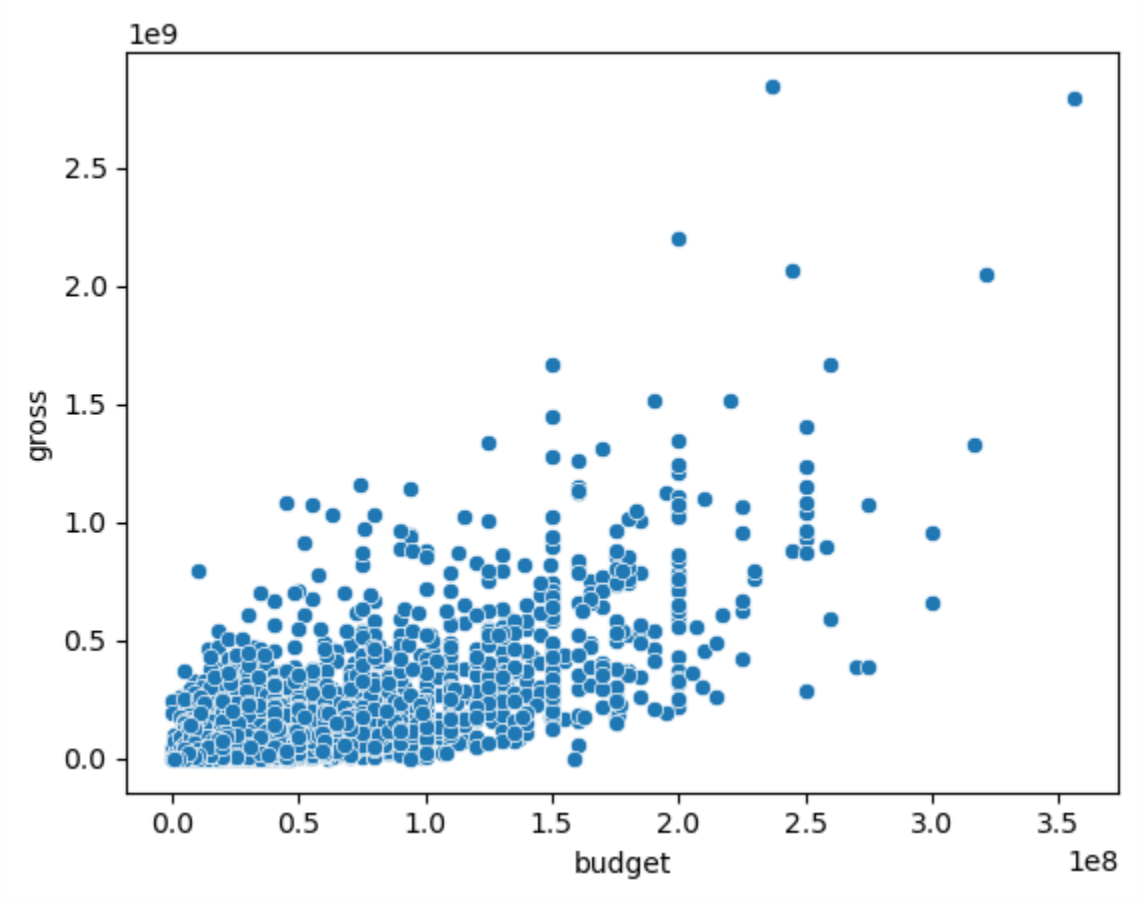
Step 2: Modeling and Analysis (8 hours)
Choose appropriate modeling techniques to answer the research questions.
Build and test predictive models to identify variables that have the strongest correlation with the target variable.
Develop visualizations and summaries of the results to present the findings of the analysis.

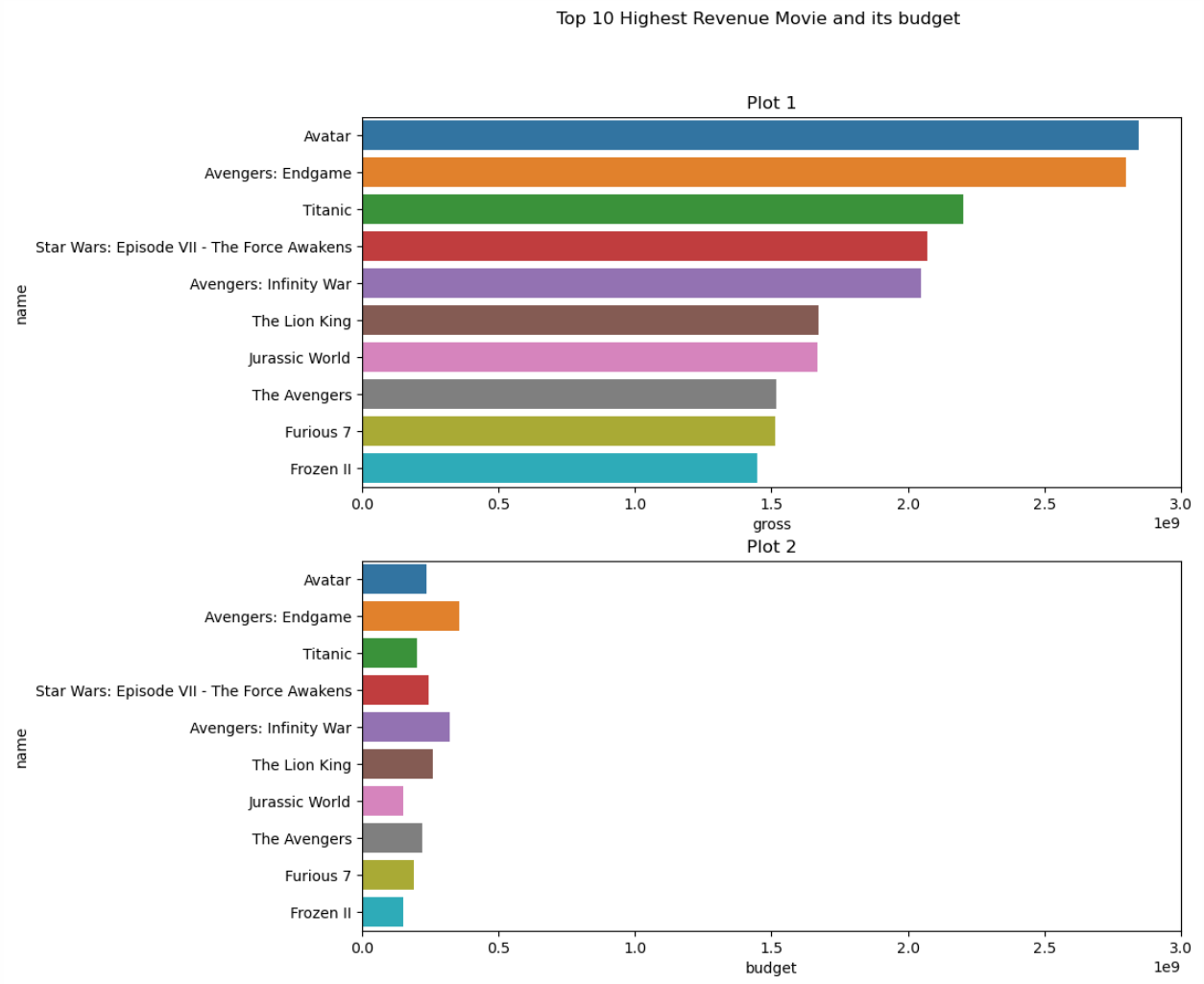
Step 3: Conclusion and representation (3 hours)
Summarize the results of the analysis and draw conclusions from the data.

Evaluate the effectiveness of the data visualization and models.
Create an engaging and informative representation.

Results:

Q1:
First I created a scatter plot to show the correlation between budget and net profit. I decided to start by looking at the 10 most successful movies. I created two cross-sectional bar charts to compare the difference between the budgets of the 10 highest-grossing films and their revenues, and could find that it was still not possible to determine the range of budgets that would affect profits in some cases. This also means that budgets are not directly related to revenues.

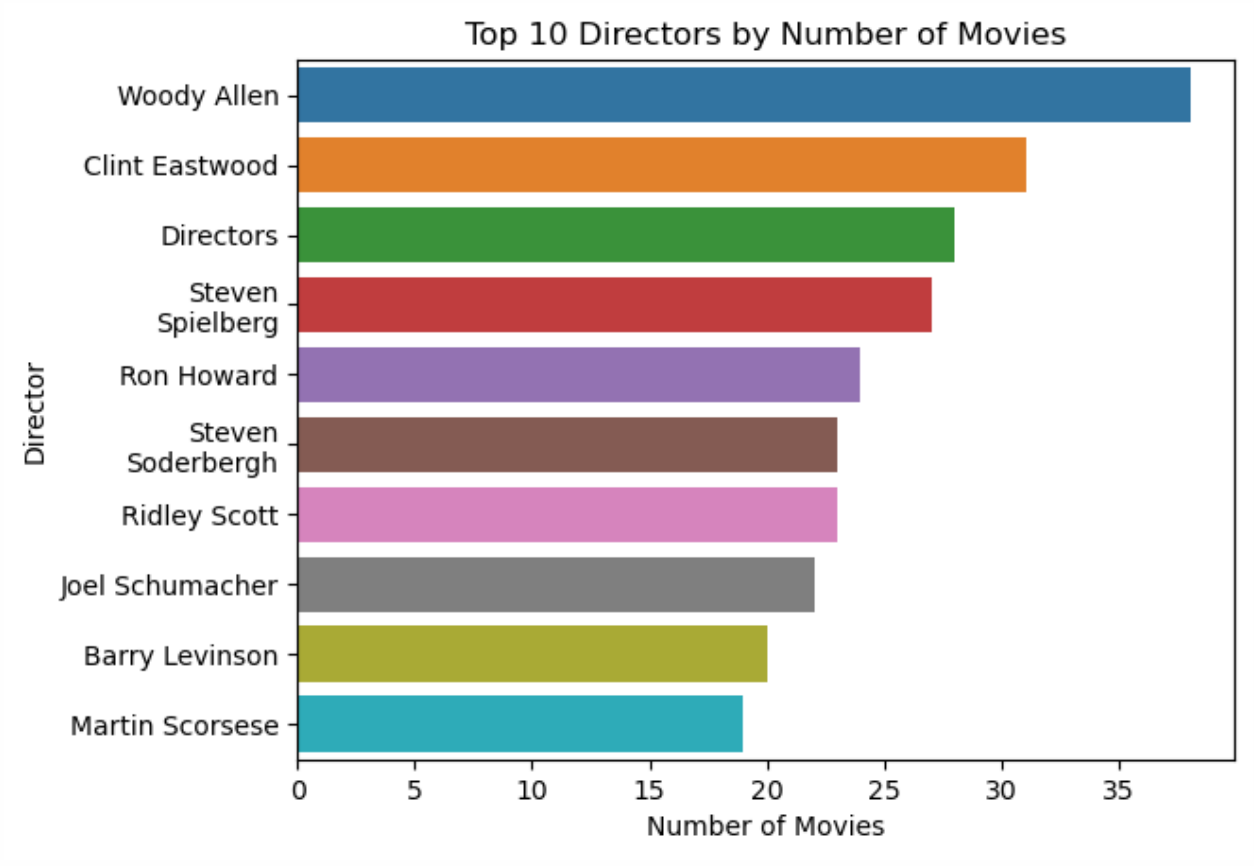


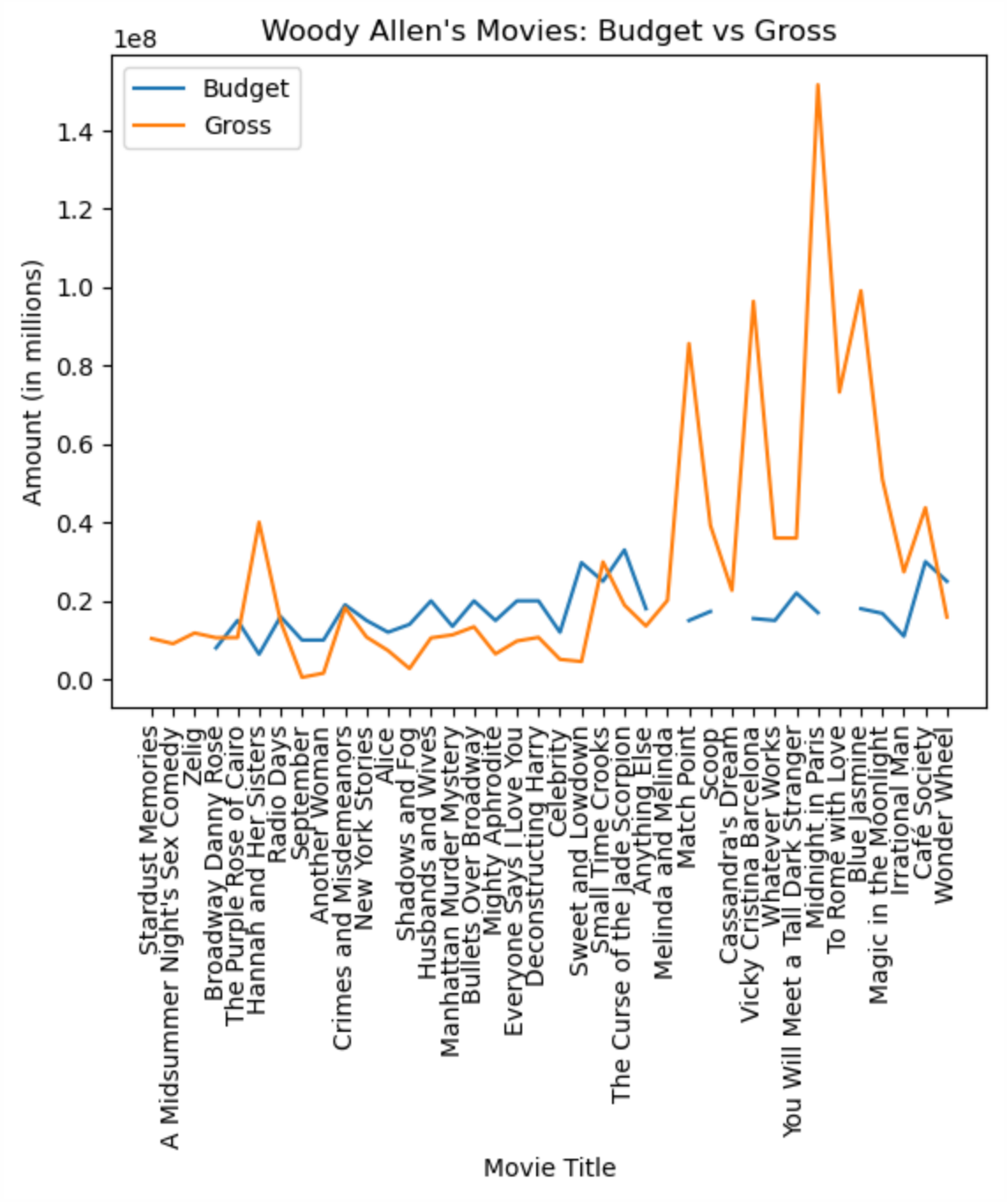


Q2:

To answer the second question, I used the number of films a director has directed as a measure of that director's reputation and performance. So I first identified the 10 directors who have directed the highest number of films, and then I created a horizontal bar chart to show it. From the bar chart, we can see that 'Woody Allen' is the "most known" director. We can take it as the most representative individual to analyze, and then I made a line chart to show the budget and revenue of all the movies directed by woody allen. Once again, we find that budgets and revenues do not correlate predictably. Some films will earn less than their budgets, while others will earn significantly more than their budgets. We can also see from the graph that the films in the second half of Woody allen's career had pretty good budgets, which means that as the director's reputation increases, the revenue generated by the film is likely to increase

as well. We can also find that a director's past performance does not seriously affect the budget given to them by the producers.

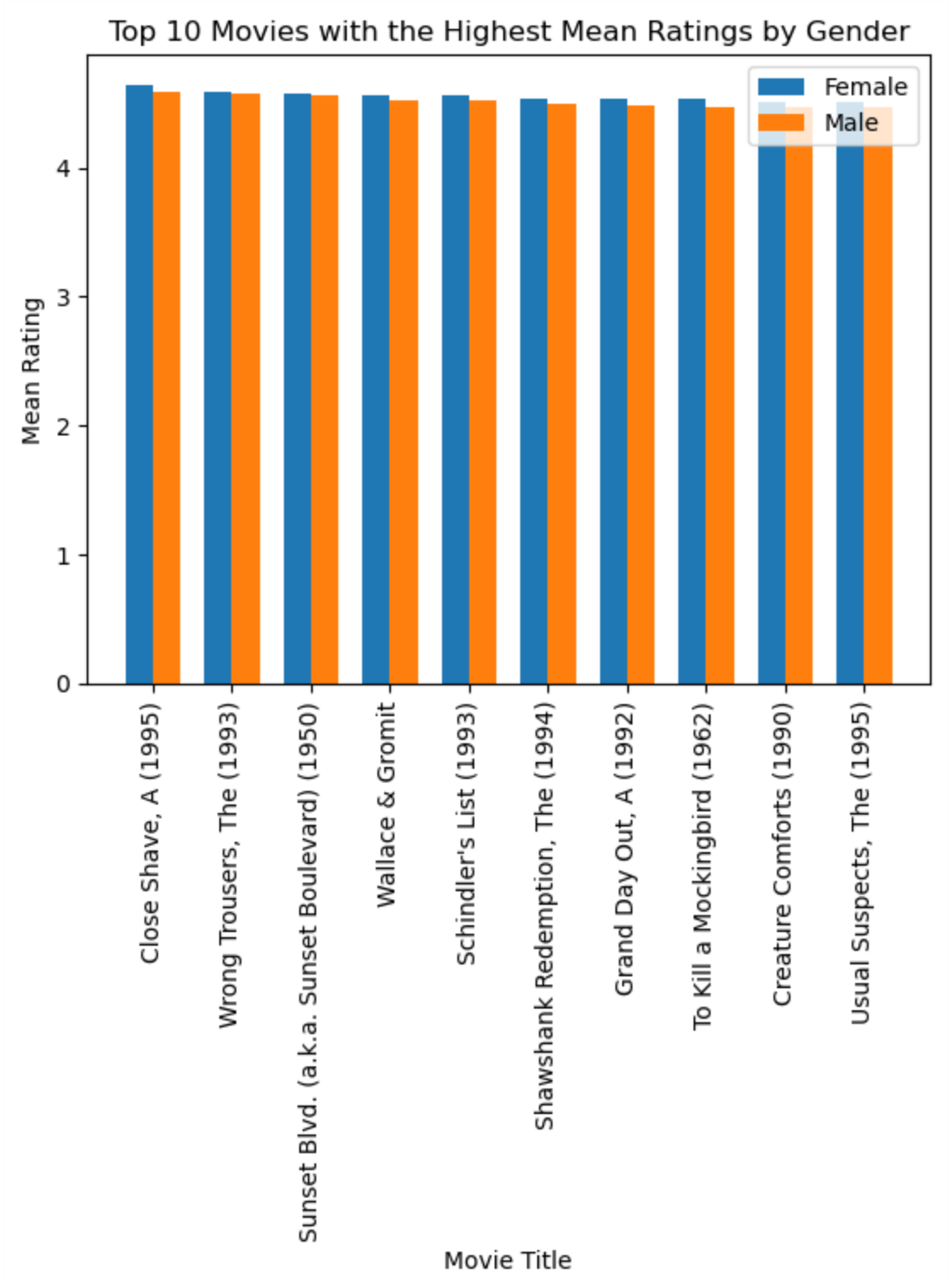


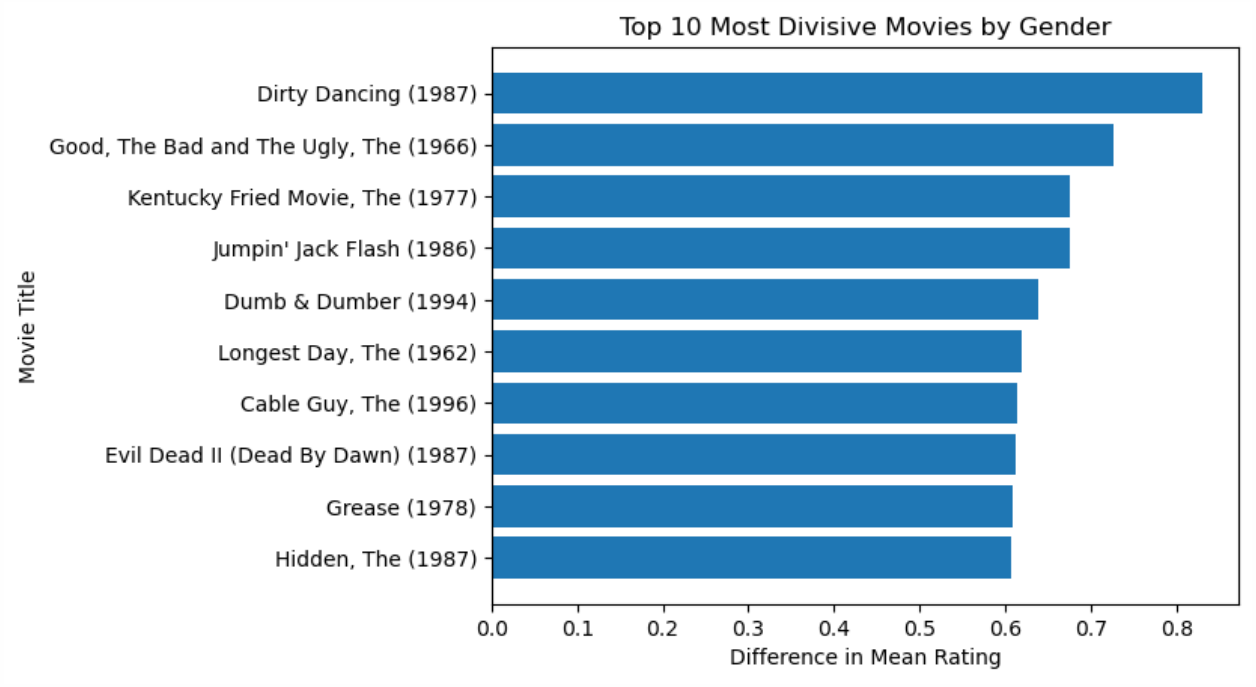


Q3:

To answer the third question, I merged three different datasets. One contains user information, another contains movie scores, and the last one contains movie information. The first and second datasets had a common variable of user id, while the second and third datasets had a common variable of movie id. So around these common variables, I first merged datasets 1, 2, and then 2, 3. Then I calculated which

movies were rated the highest among women, and which were rated the highest among men. Then I created a bar chart to show the 10 movies that were very highly rated among both men and women. Finally, a cross-sectional bar chart was created to show which movies were most highly rated by both men and women.


Top 10 Movies with the Highest Mean Ratings by Gender
Female
Male
Mean Rating
0
1
2
3
4
Close Shave, A (1995)
Wrong Trousers, The (1993)
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)
Wallace & Gromit
Schindler's List (1993)
Shawshank Redemption, The (1994)
Grand Day Out, A (1992)
To Kill a Mockingbird (1962)
Creature Comforts (1990)
Usual Suspects, The (1995)
Movie Title



Q4:
For the fourth problem, I performed a sentiment analysis using machine learning techniques with a dataset containing only reviews and sentiments. I then divided the dataset into a training set and a test set in a ratio of eight to two. The final mold trained using LogisticRegression has an accuracy of 89%. Although this is not a very high accuracy rate, it does provide some useful information in one way or another.

Impact and Limitations:
I think many people would benefit from my analysis, which can help film investors make better judgments. As well as a more credible analysis for actors or directors, it will give them a better understanding of the current state of the industry they are in. This allows them to decide what type of film is more likely to be successful next.

Challenge Goals:
The first challenge goal was within my expectation, but the second challenge goal was harder than I expected. This is because it is difficult to handle long strings, and LogisticRegression sometimes fails to run because it cannot handle too much data.

Work Plan Evaluation:

The work plan was a little different from the reality because I didn't manage my time well.

Testing:
I used assert statements to test the results of my code to make sure it was outputting the correct content. That is, the image. For the machine learning part, I didn't use assert statements to test, I printed out the results of classification_report().

Collaboration:
I did not work with anyone other than team members and faculty. I just did some google searches as well as youtube videos to learn.