Introduction

The purpose of this report is to help SussexBudgetProductions choose the genre that will optimize their future film's profitability. The business is investigating if a romantic or horror movie would do better after the recent financial failure of its comedy-action-thriller. We hope to ascertain which genre has the greatest potential for financial success by examining IMDb data to comprehend important elements such as IMDb scores, genre popularity, and revenue indicators like gross and budget.

The dataset analysis involves several steps: data cleaning and transformation, exploratory data analysis (EDA), statistical testing, and visualization. By comparing the Romantic and Horror genres, we aim to provide a recommendation for the genre with the highest revenue potential, thereby guiding SussexBudgetProductions in their strategic decision.

Methods

The analysis methodology comprises data cleaning, transformation, and statistical testing, ensuring that our results are accurate and replicable.

1. **Data Exploration and Missing Value Handling**
    - **Identifying Missing Values**: Initially, we assessed each column for missing values. Numerical columns had missing entries, which could impact revenue per capita calculations. These columns are essential for calculating profitability, so we chose to impute missing values for these columns with the median. I used median since it is more robust to outliers.
    - **Categorical Columns**: For categorical columns, we filled missing values with the mode, which represents the most frequently occurring value.
    - **Leaving NULL Values in Specific Cases**: In certain cases, such as Facebook likes data, we chose not to impute missing values. Instead, we normalized Facebook likes data for the years where the movie was released.

2. **Population Data Integration with get_population Function**
    - **Objective**: To offer context for revenue per capita, consider the country population in each movie's release year and the country where it was published. For example, a film with a high revenue per capita in a year with a lower country population could have a greater per-person impact.
    - **Implementation**: The get_population function retrieves population data for each movie's release year with the country it was published using the World Bank API. This function makes an HTTP request to fetch the world population for a specific year in each country. If data is unavailable, it returns None. Then by using a lambda function, we mapped our get_population function with the release_year and country
    - **Handling Missing Population Data**: If population data is unavailable for a specific year, we leave the world_population field as NULL for that entry. Then we dropped those rows.

3. **Data Transformation and Feature Engineering**
    - **Splitting Multi-Genre Entries**: We divided the genres column to examine each genre independently because some films belong to more than one. By calculating IMDb score distributions and movie numbers by genre, this step improves the accuracy of genre-based insights.

    - **Dropping Certain Columns**: Since they showed a poor correlation with revenue_per_capita and IMDb scores, the columns where facebook_likes in it were dropped from the final analyses. So, it means to not have great impact on any movie profitability.

4. **Key Analytical Steps**
    - **Hypothesis Testing**: We applied two different one-sided t-test to result our hypothesis. Firstly, we want to check if the revenue_per_capita parameter would result in a logical recommendation. Below you can see the test we applied for this purpose.

$$\mathbf{H_0}: \mu_{\text{revenue\_per\_capita, Romance}} - \mu_{\text{revenue\_per\_capita, Horror}} = 0$$

$$\mathbf{H_1}: \mu_{\text{revenue\_per\_capita, Romance}} - \mu_{\text{revenue\_per\_capita, Horror}} > 0$$

    - This test result in that choosing Romance would not significantly profitable. That is why we applied a second test on imdb_score.

$$\mathbf{H_0}: \mu_{\text{imdbScore, Romance}} - \mu_{\text{imdbScore, Horror}} = 0$$

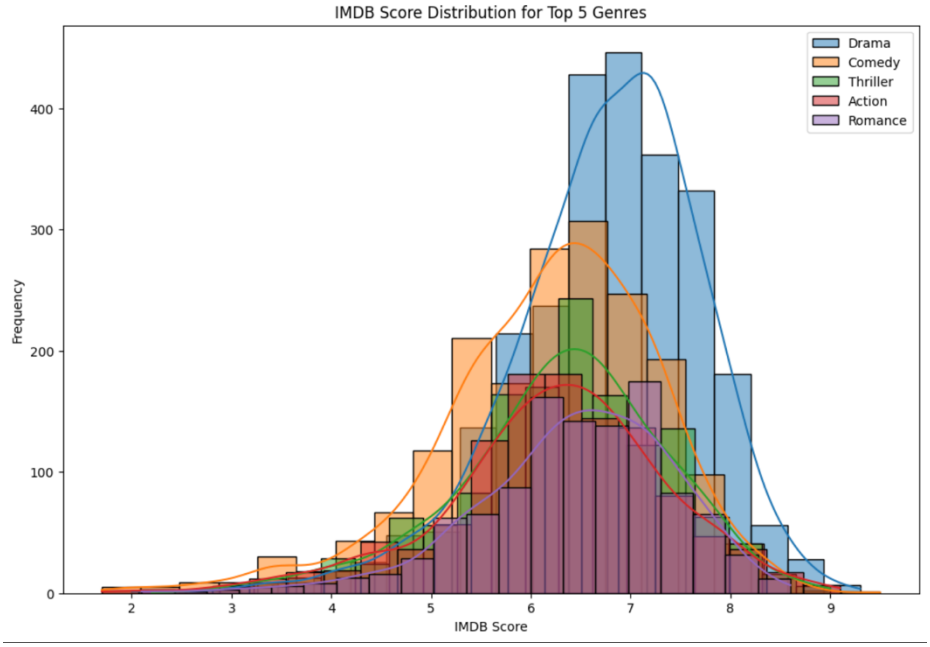
$$\mathbf{H_1}: \mu_{\text{imdbScore, Romance}} - \mu_{\text{imdbScore, Horror}} > 0$$

Results

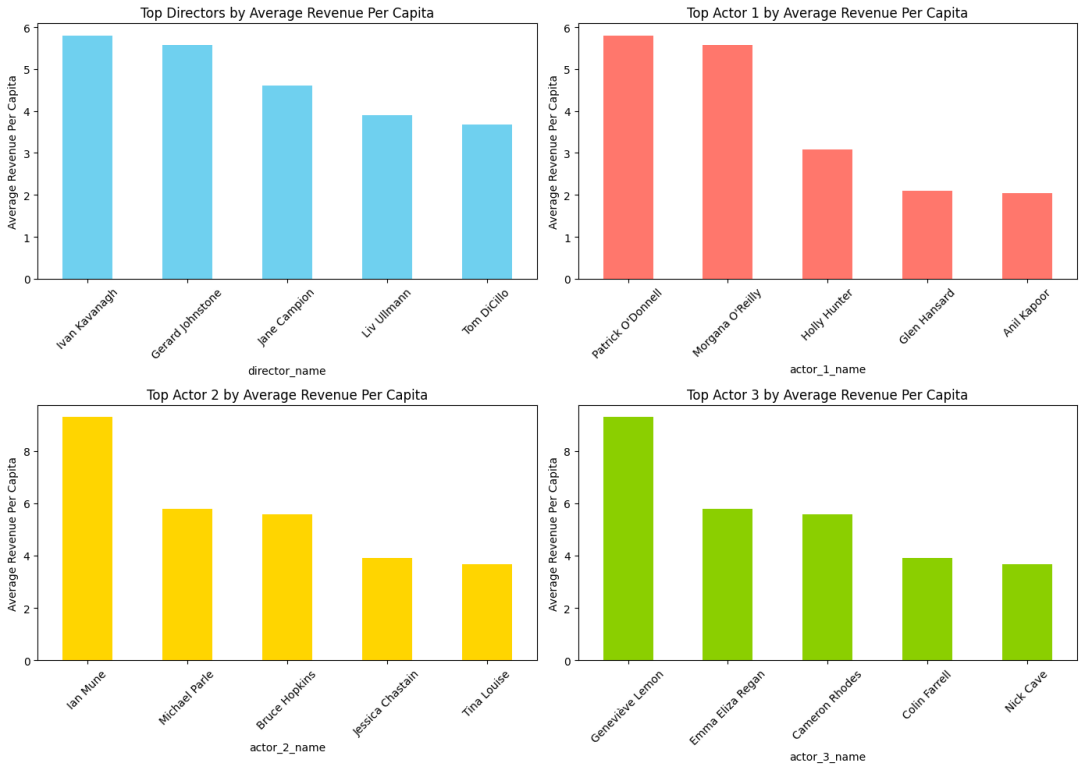
1. **Top 10 Movies**: Based on IMDb scores, the top 10 films encompass various genres, underscoring that high-quality films are not confined to one specificcategory. This diversity indicates that genre choice is essential.

	movie_title	imdb_score	genres
2765	Towering Inferno	9.5	Comedy
1937	The Shawshank Redemption	9.3	Crime|Drama
3466	The Godfather	9.2	Crime|Drama
4409	Kickboxer: Vengeance	9.1	Action
2824	Dekalog	9.1	Drama
3207	Dekalog	9.1	Drama
66	The Dark Knight	9.0	Action|Crime|Drama|Thriller
2837	The Godfather: Part II	9.0	Crime|Drama
3481	Fargo	9.0	Crime|Drama|Thriller
339	The Lord of the Rings: The Return of the King	8.9	Action|Adventure|Drama|Fantasy

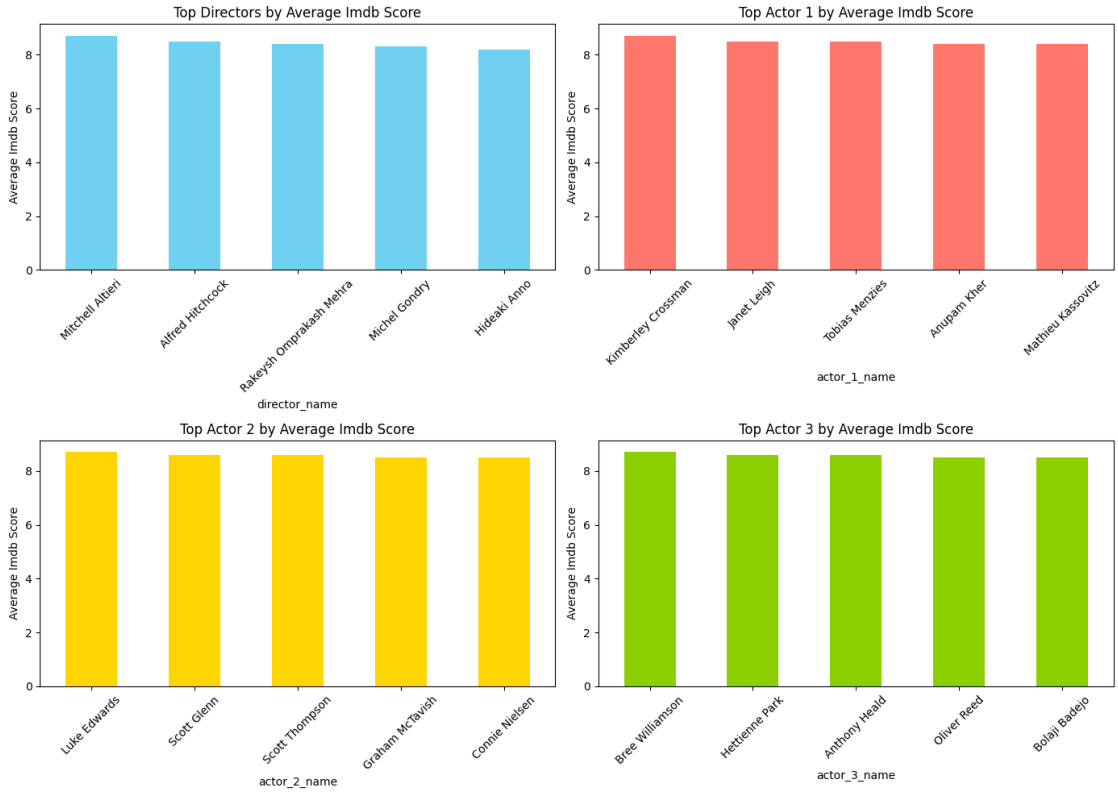
2. **Top 5 Genres**: The five most represented genres in our dataset were Action, Romance, Comedy, Horror, and Thriller. As Romance listed in top 5 genres according to IMDb Score parameter, actually we can understand that it might be profitable than Horror movies.



3. **Top Directors And Actors:**
   - We conduct 2 different methods to find the best 5 directors and actors. The first method's essential parameter was revenue_per_capita. It results is as follows:



- The second method was based on IMDb Score each director and actors got. Here is the results:



4. **Hypothesis Test Findings**:
   - Our t-test results showed that Romance movies does not have a significant higher revenue_per_capita than Horror Movies. In fact, if our critical parameter would be revenue_per_capita, then choosing Horror movies would most probably more profitable.

```
T-Statistic: -1.3393929310597197
P-Value: 0.9096347540100786
Fail to reject the null hypothesis (H0): No significant evidence that Romance movies have higher mean revenue_per_capita than Horror movies.
```

   - The second t-test results showed that Romance movies does have a significant higher IMDb Score than Horror movies. In our case, if we assume IMDb Score is the one of the criteria for the profit than we would like to select Romance movies.

```
T-Statistic ( IMDb Score ): 10.673456509513956
P-Value ( IMDb Score ): 1.5816670359309433e-25
Reject the null hypothesis (H0): Romance movies have a significantly higher mean imdb_score than Horror movies.
```

Conclusion

Based on the analysis, a Romance genre film is recommended for SussexBudgetProductions. Romance movies tend to achieve higher IMDb scores, which align with the company's goal of producing a profitable film. To maximize the chances of success, selecting a director and cast with proven experience in the Romance genre is advised. This selection could help capitalize on the genre's appeal and enhance audience engagement. Here is the top directors, and actors for the Romance movies.