In this study, by examining the historical performance data of the films, SussexBudgetProductions company's next project will determine the film genre that is likely to make a profit(only horror or romance films are permitted) and the names of the actor and director that may be recommended for this film. Because a high IMDb score indicates that a movie is liked and of high quality, the senior leadership team believes that a high IMDb score will guarantee a high profit. For this reason, the company aims to make a movie with a high IMDb score. There are many categorical(color, duration, release_year) and numerical(imdb_score, gross, budget) columns in the dataset where the quality and financial success of a movie can be measured. After the data wrangling/cleaning steps, the average profit percentages and IMDb scores of romantic and horror movies are tested with a two-tailed t-test to decide which type of movie tends to be a high-profit return. These tests aim to determine whether one of the romantic or horror genres is more advantageous in terms of high profitability or IMDb score.

Firstly, some column names have been updated to provide more information (For example, a unit has been added to the column name: duration -> duration_minutes). After that, the spaces at the beginning and end of the data in the movie_title column have been removed. (For example, there are many gaps in the movie named ‘Happy Valley’ in the original data.) Later, all characters in the dataset were converted to lowercase letters to prevent incorrect groupings. The genres column separated from each other by the ’|' symbol is processed into the genres_updated column, based on whether each entry contains "horror," "romance," both, or neither. Duplicated data, based on director_name, movie_title, release_year, is dropped by keeping the row with maximum voted_users_count value because higher voted_users_count value represents more updated data. All NaN values found in critical columns (budget, gross_revenue, imdb_score, genres, director_name, actor_1_name) have been dropped to ensure consistency. Since the number of nulls in the remaining columns is low and have no significance to affect the analysis, the categorical columns were filled with the ‘unknown’ value; the numerical columns were filled with the median value because their skewness value is high.

| | column | skewness |
|---|---|---|
| 0 | critic_review_count | 1.455411 |
| 1 | duration_minutes | 2.347947 |
| 2 | actor_3_fb_likes | 6.623984 |
| 3 | faces_in_poster_count | 4.870509 |
| 4 | actor_2_fb_likes | 9.446795 |
| 5 | aspect_ratio | 15.963330 |

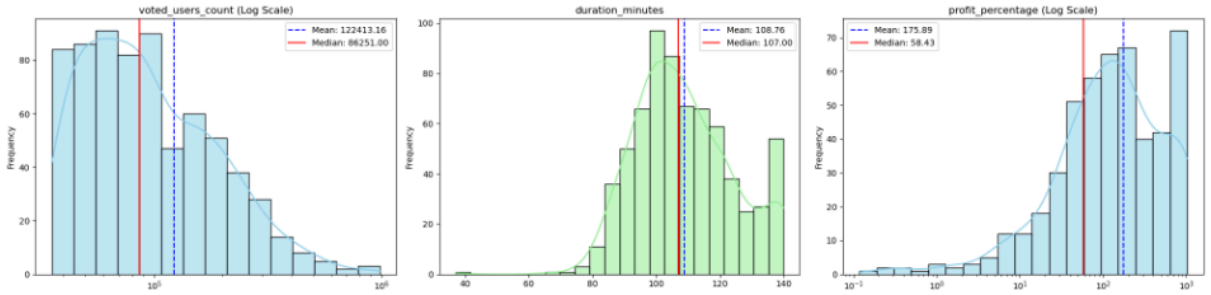
| | null_count |
|---|---|
| aspect_ratio | 72 |
| content_rating | 49 |
| plot_keywords | 31 |
| actor_3_fb_likes | 7 |
| actor_3_name | 7 |
| faces_in_poster_count | 6 |
| language | 4 |
| color | 2 |
| actor_2_name | 2 |
| actor_2_fb_likes | 2 |
| critic_review_count | 1 |
| duration_minutes | 1 |

Non-country entries in the Country column (e.g., "Official Site," "New Line") were replaced with the column's mode value. Additionally , "West Germany" values were consolidated under "Germany". Profit was calculated by subtracting the budget column from the gross_revenue column. Since the accuracy of a proportional approximation will be higher in the analyses, the profit_percentage value has also been calculated. Since only horror and

romantic movie genres will be dealt with, the data of these two genres has been transferred to a new dataframe. To make outlier detection in this dataframe, distribution graphs and some statistical values were calculated.

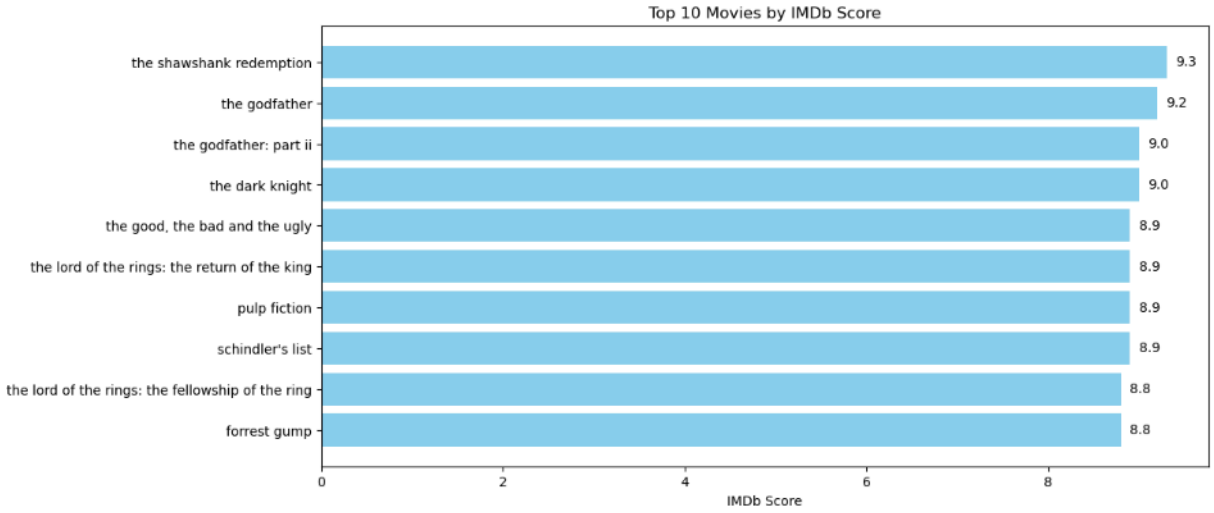
| | Mean | Median | Variance | Standard Deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| **critic_review_count** | 1.518376e+02 | 1.310000e+02 | 1.055409e+04 | 1.027331e+02 | 1.256520 | 2.156085 |
| **duration_minutes** | 1.072469e+02 | 1.030000e+02 | 3.738575e+02 | 1.933539e+01 | 1.899445 | 7.680080 |
| **director_fb_likes** | 4.692212e+02 | 4.900000e+01 | 4.607767e+06 | 2.146571e+03 | 6.560385 | 45.627207 |
| **actor_3_fb_likes** | 6.424114e+02 | 3.980000e+02 | 2.500868e+06 | 1.581413e+03 | 7.337657 | 59.344273 |
| **actor_1_fb_likes** | 6.331754e+03 | 1.000000e+03 | 1.194424e+08 | 1.092897e+04 | 5.271646 | 55.196193 |
| **gross_revenue** | 4.039331e+07 | 2.547297e+07 | 2.802571e+15 | 5.293932e+07 | 3.681974 | 24.389022 |
| **voted_users_count** | 7.565609e+04 | 4.386700e+04 | 9.424210e+09 | 9.707837e+04 | 3.328830 | 17.426320 |
| **cast_total_fb_likes** | 9.465410e+03 | 3.299000e+03 | 2.456422e+08 | 1.567298e+04 | 7.268038 | 108.528013 |
| **faces_in_poster_count** | 1.318367e+00 | 1.000000e+00 | 4.700847e+00 | 2.168144e+00 | 7.340073 | 115.672577 |
| **user_review_count** | 2.930824e+02 | 1.970000e+02 | 1.055701e+05 | 3.249156e+02 | 3.514277 | 18.575818 |
| **budget** | 3.962754e+07 | 1.800000e+07 | 1.273851e+17 | 3.569104e+08 | 32.762205 | 1110.867732 |
| **release_year** | 2.002358e+03 | 2.004000e+03 | 1.069766e+02 | 1.034295e+01 | -2.444676 | 10.222758 |
| **actor_2_fb_likes** | 1.643972e+03 | 6.430000e+02 | 2.600399e+07 | 5.099411e+03 | 16.354625 | 407.921012 |
| **imdb_score** | 6.273796e+00 | 6.400000e+00 | 1.011576e+00 | 1.005771e+00 | -0.644250 | 0.787946 |
| **aspect_ratio** | 2.086571e+00 | 2.200000e+00 | 7.488497e-02 | 2.736512e-01 | -0.318699 | -1.116312 |
| **movie_fb_likes** | 6.789074e+03 | 8.800000e+01 | 2.822997e+08 | 1.680178e+04 | 4.392179 | 24.445679 |
| **profit** | 7.657761e+05 | 2.817771e+06 | 1.293508e+17 | 3.596537e+08 | -32.324438 | 1090.991023 |
| **profit_percentage** | 1.085627e+03 | 2.365311e+01 | 4.705798e+08 | 2.169285e+04 | 30.730864 | 994.711467 |

Because the skewness values are lower for critic_review_count, duration_minutes, faces_in_poster_count, user_review_count, and profit_percentage columns, only a 5% outlier limit has been applied. Because the skewness is higher for the voted_users_count column, a 25% lower band has been applied. Thus, rows with a small number of comments but with a high IMDb score will not be misinterpreted during the analysis processes. The corresponding threshold has been tried multiple times and the value that will obtain the ideal distribution graph has been found.

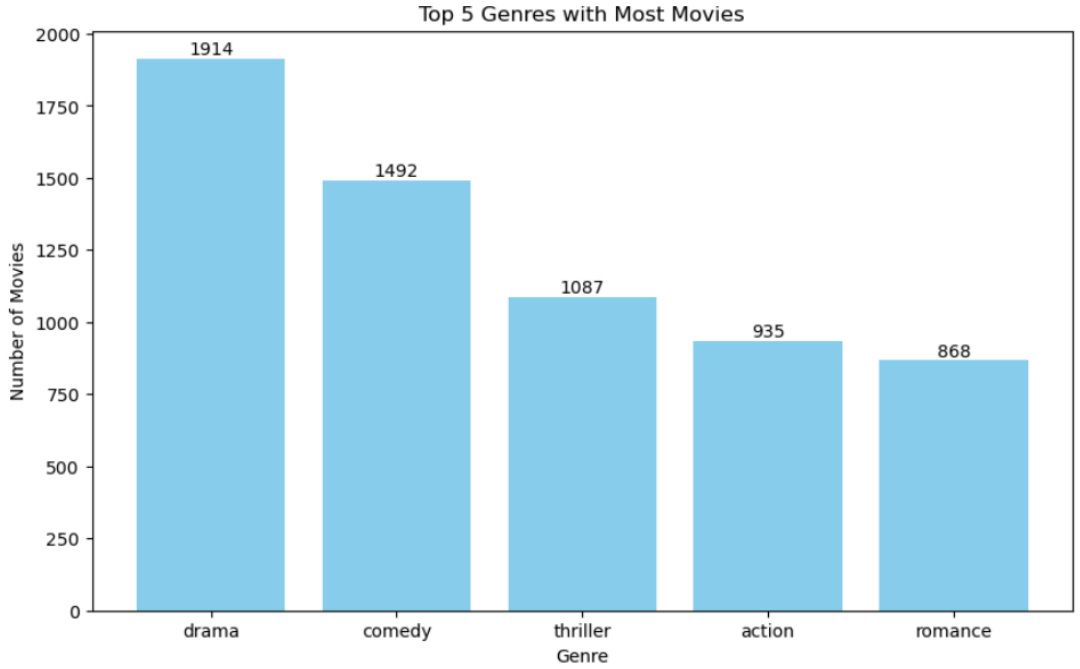



As a result of the applied two-tailed t-test, horror films tend to have a higher profit average. At the same time, romantic movies also tend to have higher IMDb scores.

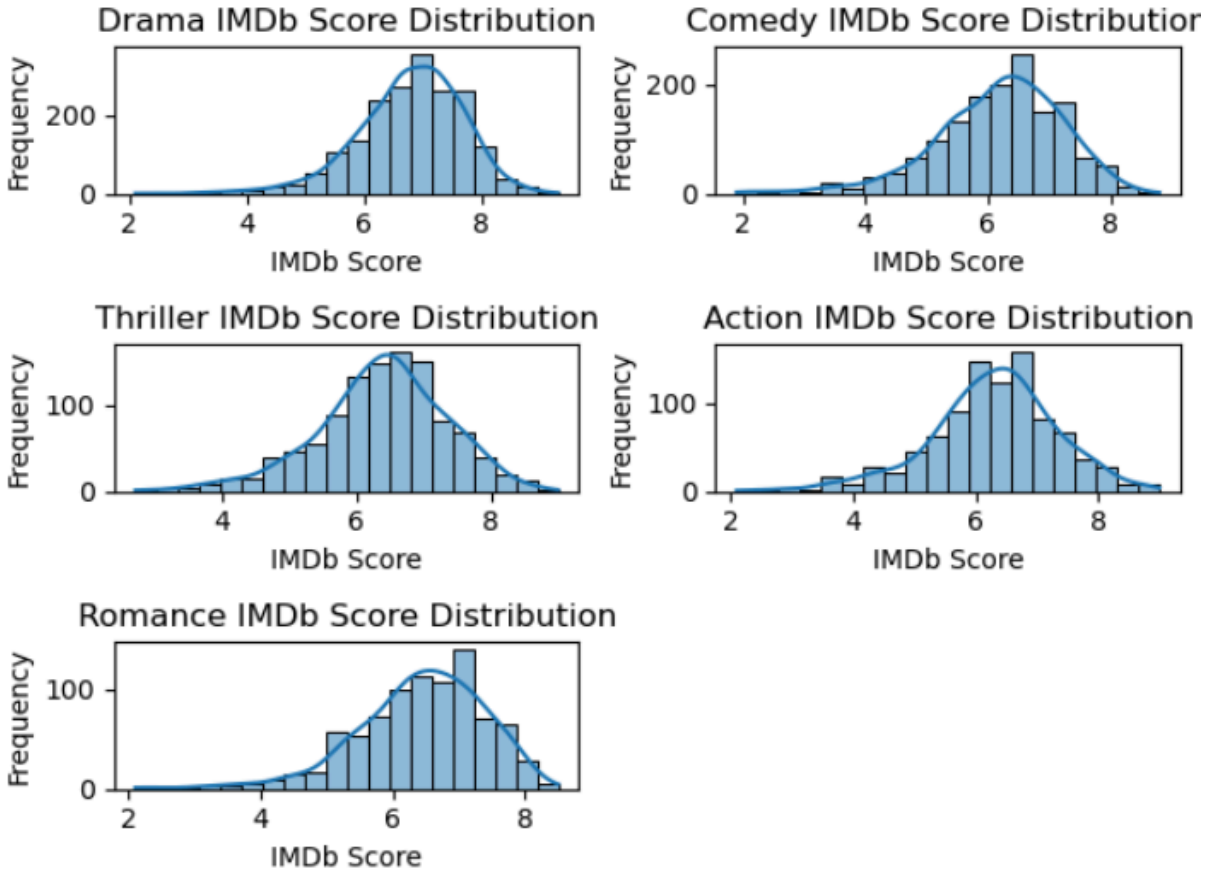
Top 10 highest IMDb scores include notable films like *The Shawshank Redemption* and *The Godfather*. This is the relationship between the IMDb indicator and movie popularity.



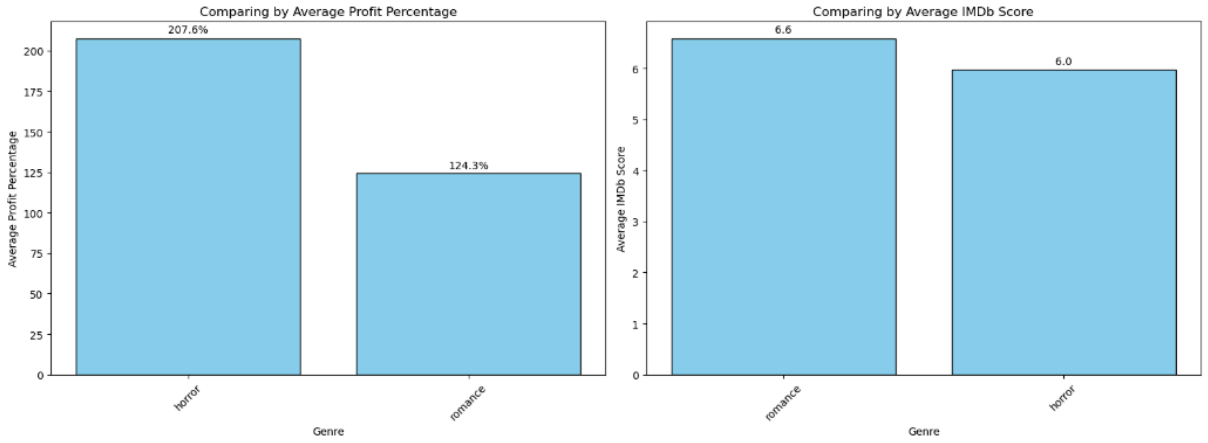

At the same time, when the top 5 genres with the most films are listed, it seems that the drama genre is the most common.




Based on the average IMDb scores of the genres in the graph, the relationship between the genres and IMDb scores can also be established. Additionally, below it can be seen the IMDB score distribution of the top 5 genres:

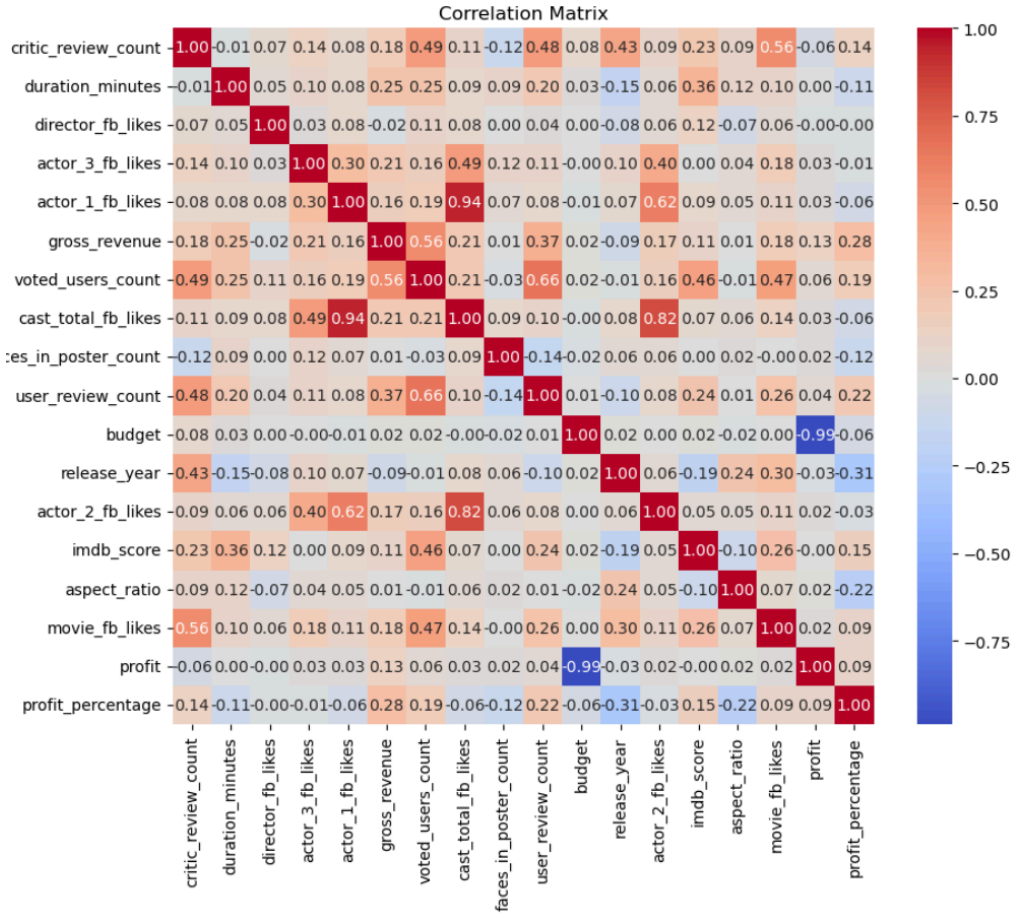



In the graph comparing horror and romantic movie genres based on the average IMDb score and the average profit ratio, horror movies tend to have a higher average profit value, while romantic movies also tend to have a higher IMDb score.




These graphs also confirm the two-tailed t-test result.
First hypothesis: “There is a significant difference between the average profit percentages of romantic and horror genres.”
Result: Because t-statistic value is negative and p_value smaller than 0.05, horror films tend to make more profit.
Second hypothesis: “There is a significant difference between the romantic and horror genres in terms of IMDB score.”
Result: Because t-statistic value is positive and p_value smaller than 0.05, romance films tend to have higher IMDb score.

Because the company's main goal is profitability, it would be appropriate to make a horror movie. In order for the actor and director to be selected to increase the high IMDb score and profit with the company's goals, the average IMDb score and the average profit ratio in films in which he takes part in the horror genre have been calculated. Therefore, horror movies were grouped by director and actor to calculate average profit percentage and IMDb score. The reason why the IMDb score is taken into consideration is that it correlates with the profit percentage and the company aims to generate profitable films with a high IMDb score. It was found appropriate to choose the actor and director with the highest IMDb score and a profit rate higher than the average profit rate of horror films. The top director and actor with above-average profit percentages were selected for the film.




(

| director_name | profit_percentage | imdb_score |
|---|---|---|
| jonathan demme | 588.036842 | 8.6 |
| alfred hitchcock | 1030.684389 | 8.5 |
| ridley scott | 617.272727 | 8.5 |
| william friedkin | 1030.684389 | 8.0 |
| edgar wright | 236.609700 | 8.0, |

| actor_1_name | profit_percentage | imdb_score |
|---|---|---|
| janet leigh | 1030.684389 | 8.5 |
| tom skerritt | 617.272727 | 8.5 |
| ellen burstyn | 1030.684389 | 8.0 |
| peter serafinowicz | 236.609700 | 8.0 |
| shane black | 298.236987 | 7.8) |

Finally, horror and romance genres are taken into account, the genre that tends to provide the most profit rate according to test results and graphic outputs is the horror genre. Therefore, the next genre of film should be horror and Jonathan Demme and Janet Leigh should be selected as a director and actor respectively.