**Title: Movie Success Insights: Statistics, Popularity and Timing Considerations**

**Executive Summary**

In the ever-evolving landscape of the film industry, our comprehensive analysis delves into key factors shaping the success of movies. Through a meticulous exploration, our research unveils intricate insights spanning the evolution of ratings, discerning audience genre preferences, elucidating the profound impact of actors and directors, analysing the delicate orchestration of budgets, and illuminating the strategic interplay of timing. These data-driven revelations are strategically positioned to empower decision-makers within the film industry, providing invaluable guidance to navigate the ever-shifting landscape. By leveraging these insights, industry leaders can contribute to the creation of enduring, impactful cinematic experiences, ensuring a sustained influence within the dynamic realm of film production.

**Data Sources**

The datasets comprised two files: 'OMDBmoviedata.csv' (retrieved using an API GET request at <https://www.omdbapi.com/>) and 'movies\_data.csv' (sourced from <https://github.com/danielgrijalva/movie-stats>). The datasets encompassed key attributes such as 'Movie\_Name,' 'Rating,' 'Genre,' 'Year,' 'Rating\_Score,' 'Votes,' 'Director,' 'Writer,' 'Star,' 'Country,' 'Budget,' 'Gross,' 'Company,' 'Runtime,' 'Release\_date,' 'Production,' 'Poster,' 'Type,' 'Ratings,' 'Website,' 'totalSeasons,' 'Error,' and 'Response.' The movie\_data.csv dataset was combined with 'OMDBmoviedata.csv' for the analysis due to the inclusion of the 'budget' column required for the analysis.

To ensure data quality and relevance, we conducted thorough data cleaning, addressed missing values, and formatted specific columns appropriately. Employing techniques such as grouping, aggregation, summary statistics, correlation, regression, independent sample t-tests, ranges, and plots, we adhered to a structured methodology. This approach ensured that the data underwent comprehensive processing, rendering it well-prepared and suitable for subsequent analysis.

**Data Processing and Pandas Integration**

The Python programming language (version 3.10.13 packaged by Anaconda, Inc.) alongside version 2.0.3 of the Pandas library is utilised within Jupyter Notebook (version 5.3.0) for data processing and analysis. Leveraging Pandas efficiently facilitated the organization and manipulation the information into a structured tabular format, primarily using the Pandas DataFrame as the primary data structure. The DataFrame facilitated data cleaning, handling of missing values, and overall data preparation for a meaningful and insightful analysis. Furthermore, the codes were generated using pandas, numpy, matplotlib, seaborn and scipy.stats documentations (Pandas Documentation, 2023).

**Research Questions**

* OMDB Movie Ratings and Trends:
  + *Question:* How have movie ratings evolved over the years, and are there discernible trends within different genres or directors?
* Genre Analysis and Popularity:
  + *Question:* Which movie genres reign supreme in popularity, and is there a correlation between a movie's genre and its commercial success?
* Director/Actor Influence on Movie Success:
  + *Question:* Do certain actors or directors wield a significant influence on a movie's success or ratings?
* Movie Length and Audience Preference:
  + *Question:* How does the duration of a movie impact audience reception and ratings? Are longer or shorter movies more favorably received?
* Box Office Performance and Budget Analysis:
  + *Question:* Is there a relationship between a movie's budget, box office performance, ratings, and how timing influences these dynamics?

**Hypothesis**

**1. OMDB Movie Ratings and Trends**

1. Alternate Hypothesis: The times series is stationary, so ratings do not vary over time and fluctuate around a constant level.
2. Null Hypothesis: The time series is non-stationary, so ratings vary over time through trends and seasonality.

**2. Genre Analysis and Popularity**

1. Alternate Hypothesis: The greater the popularity of a movie genre, the more successful it will be.
2. Null Hypothesis: The popularity of a movie genre does not have a significant positive effect on its success.

**3. Director/Actor Influence on Movie Success**

1. Alternate Hypothesis: Directors or actors wield a significant influence on a movie's success.
2. Null Hypothesis: Directors or actors do not wield a significant influence on a movie's success.

**4. Movie Length and Audience Preference:**

**5.** **Box Office Performance and Budget Analysis**

1. Alternate Hypothesis: there a relationship between a movie's budget, box office performance, and ratings.
2. Null Hypothesis: there is no relationship between a movie's budget, box office performance, and ratings.

**Key Findings and Analysis**

1. **OMDB Movie Ratings and Trends**

**Explore the rating scores between IMDB and Metascore over the years:**

1. The ratings given by Metascore and IMDB are provided in two different ways. The Metascore is a weighted average of reviews from critics and publications for a given movie along with user scores. IMDB, meanwhile, is open to anyone with an IMDB account. IMDB movie scores can vary wildly depending on the user, though the website does try and ensure extreme scores don't adversely affect the rating. Since it doesn't rely on critical evaluation, the score it's far more of an indicator of popularity than it is actual quality.
2. Starting with the IMDb line, we can see that most ratings are between 6 and 8.5. The highest point of the lines shows in the earlier years, with the line gradually settling between 6.5 and 7 from 1978 onwards before suddenly spiking back up from 2020 and onwards.
3. The Metascore line is distributed more widely between 5.5 and 10. It follows a similar pattern to the IMDB line, though it rises higher than the IMDB line in the earlier years. It then falls below the IMDB line from 1978 onwards, before spiking back up in 2020.
4. The p-value for both Metascore and IMDB is greater than or equal to 0.05, so the null hypothesis would not be rejected and the alternative hypothesis would be rejected to show that ratings do vary through trends, and it would be affected by other factors.

**Explore the number of votes given to movies on IMDB over the years:**

1. When looking along the line, there are noticeable peaks throughout certain years where there are a large number of voters on average. With IMDB having started as a company in 1990, it may be that these older movies have had more time to accumulate ratings. It is also the case that classic films continue to be highly regarded by people, with classics like Spartacus/Psycho in 1960, Grease in 1978 and The Lion King in 1994. While taking these peaks into account, it can also be seen that the line is generally tending positively, getting slowly higher over the years before a huge spike in 2020 which can also be seen with the rating scores.

**Explore the distribution of the highest grossing movie genres over the years:**

1. Near to the start of the figure, the lines for the genres were at relatively similar levels showing a wide range in high grossing films. This then quickly spreads out, with Action, Comedy, Drama and Adventure increasing dramatically which shows that these genres have become a big draw for audiences and also highlights the growing number of movies that are coming out year upon year.
2. **Genre Analysis and Popularity**

**Popularity of Genres based on Average Votes and Rating Score**

1. **Average Votes:**  'Mystery' movies received the highest votes on average with over 250,000 votes. The next most popular genre is 'Family' movies with approximately 170,000 votes.
2. **Rating Score:** The movie genre to produce the highest average rating was ‘Biography’. However, there is not a great deal of variation in the average ratings of each genre; majority of ratings fall between 6 and 7. The small variance indicates that the rating scores tend to be very close to the mean, and to each other, therefore the movie ratings data does not provide substantial evidence to explore the hypothesis.

**Most successful Genre based on Gross Earnings**

The highest grossing genre of movies with over $500 million profit are 'Family' movies, producing almost double than 'Animation' movies.

**Linear Regression Analysis**

1. **Average Votes vs. Average Gross:** The positive correlation (0.51) between the Average Votes and Average Gross per Genre suggests that there is a moderate tendency for genres to be more successful if they have greater votes. However , since it is only a moderate correlation, it does not indicate causation.
2. **Average Rating Score vs. Average Gross:** There is a weak positive correlation (0.32) between the average ratings and gross earnings per Genre. This suggests that the success of a movie genre is not greatly affected by its rating.
3. **Actor/Director Influence**

**Top Directors and Actors Based on Average Ratings:**

1. **Directors:** Roberto Benigni, Tony Kaye, and Nadine Labaki lead in average ratings, indicating consistent high-quality filmmaking.
2. **Actors:** Mark Hamill and Philippe Noiret emerge as top-rated actors, showcasing their impact on the audience's perception of movie quality.

**Total Gross Earnings by Directors and Actors:**

1. **Directors:** Steven Spielberg leads in total gross earnings, followed by Anthony Russo and Peter Jackson. These directors consistently deliver movies that attract significant audiences and generate substantial revenue.
2. **Actors:** Robert Downey Jr. tops the list in total gross earnings, followed by Tom Hanks and Tom Cruise. These demonstrate their ability to draw audiences and contribute to a movie's financial success with impressive cumulative worldwide box office earnings.

**Average Votes:**

1. **Directors:**

The average votes received by directors, with Christopher Nolan leading the pack, indicate their influence on audience engagement. Directors who can capture the audience's attention and appreciation contribute significantly to a movie's overall success.

1. **Actors:**

Actors like Mark Hamill, Ben Burtt, and Alexandre Rodrigues garner high average votes, emphasizing their role in creating movies that resonate with audiences and receive positive feedback.

**Linear Regression Analysis:**

1. **Director's Rating vs. Actor's Rating:** A positive linear relationship is observed, suggesting that higher-rated directors tend to work with higher-rated actors.
2. **Total Gross by Director vs. Total Gross by Actor:** The weak correlation (R=0.08) implies that directorial success does not strongly predict actor-based financial success.
3. **Average Votes by Director vs. Average Votes by Actor:** A moderate correlation (R=0.38) indicates a positive relationship between the average votes received by directors and actors.

**Two-Sample T-Tests:**

1. **Average Votes Comparison:** The t-test fails to reject the null hypothesis, suggesting no significant difference in average votes between directors and actors, as indicated by the high p-value (p = 0.9405).
2. **Total Gross Comparison:** The t-test fails to reject the null hypothesis, suggesting no significant difference in total gross earnings between directors and actors (p = 0.1865).
3. **Average Votes by Director vs. Actor:** The analysis fails to reject the null hypothesis, indicating no significant difference in average votes between directors and actors (p = 0.9405).
4. **Movie Length and Audience Preference**

**Top Ten Movies with the highest votes**

Based on the top ten movies according to number of votes per movie, it suggests that there might be preferences. as shown in fig4.1, number 9 of the top ten movies, The lord of the ring: The return of the king, with a run time of 201 minutes, had a total vote of 1,700,000. While Interstellar, the last of the top ten, with a runtime of 169 minutes, had a total vote of 1,600,000. The top two movies, Darknight and Shawshank Redemption, based on votes of 2,400,00 each, have lower runtimes of 150 and 125 minutes respectively.

**Relationship Between Runtime and Viewers Preference (Hypothesis testing)**

The correlation between votes and run time is 0.30, showing a postive relationship, however weak. Correlation chart plotted, using all 76668 observations, in fig4.3 shows that there is little to no linear relationship between how viewers feel about a movie, and its runtime. Hypothesis test done using a sample size of 500 observation returned a p-value of 2.3, which is significantly higher than 0.05 (assumed significance level). Therefore we cannot reject the null hypothesis of non-significance.

**5. Box Office Performance and Budget Analysis**

**Budget vs. Box Office Earnings:**

1. An r-value of 0.5476 indicates a moderate positive linear relationship between the budget and gross of movies. As the budget increases, there tends to be a moderate increase in gross, but it's not as strong as a higher correlation would suggest.
2. The moderate correlation suggests that while there is a tendency for movies with higher budgets to have higher gross earnings, there is significant variability. In other words, there are instances where movies with lower budgets might still perform well in terms of gross earnings, and vice versa.

**Budget vs. Ratings:**

1. An r-value of 0.0051 indicates an almost non-existent linear relationship between the budget allocated for a movie and its rating. Changes in the budget don't show a significant pattern or impact on the movie ratings.
2. It suggests that the budget spent on a movie does not significantly influence or predict the rating it receives. Therefore, when aiming for a higher movie rating, other factors beyond budget allocation need to be considered and prioritized; like quality of the script, acting performances, audience preferences and so on.

**Relationship between Average Budget and Average Gross of Movies over the Years:**

1. A Pearson correlation coefficient of approximately 0.94 suggests a strong positive linear relationship between the average budget and average gross of movies over the years. The p-value associated with this correlation is extremely small (3.26e-19), indicating strong evidence against the null hypothesis that there is no correlation between the variables. This very low p-value suggests that the correlation coefficient is statistically significant.
2. This strong positive correlation over the years could reflect changes in audience preferences, production quality or other factors that have led to increased box office revenues for movies with higher budgets. While a high budget might correlate with higher gross, it also signifies a higher financial risk if the movie doesn't perform well.
3. Although there's a strong correlation, it's important to note that correlation doesn’t imply causation. Other factors, such as marketing strategies, script quality, casting, release timing, etc., also play crucial roles in a movie's success at the box office.

**Recommendations**

1. **Strategic Release Timing:** Optimize movie releases by strategically aligning them with audience preferences, seasonal trends, and competitive landscapes, based on historical data analysis.
2. **Genre-Tailored Marketing Strategies:** Tailor marketing efforts to match audience genre preferences, utilizing social media, digital platforms, and genre-specific events for targeted promotions.
3. **Investment in Directorial and Acting Talent:** Continue investing in proven and emerging directorial and acting talent, recognizing their substantial impact on critical acclaim and box office performance.
4. **Budget Allocation and Risk Mitigation:** Adopt a balanced approach to budget allocation, considering factors beyond financial investment to mitigate risks associated with higher budgets.
5. **Continuous Audience Engagement:** Prioritize continuous engagement through social media, fan communities, and feedback mechanisms to sustain interest and anticipation for upcoming projects.
6. **Strategic Partnerships:** Collaborate with successful entities in the industry, such as renowned directors, actors, or production companies, to enhance a movie's visibility and credibility.
7. **Adaptability to Industry Trends:** Embrace innovation in storytelling, production techniques, and distribution channels to stay relevant and competitive amidst evolving industry trends.
8. **Optimal storytelling:** **Optimize movie length based on storytelling needs rather than rigid norms, ensuring an immersive viewer experience aligned with narrative demands.**
9. **Continuous Data Analysis:** Establish a robust system for real-time data analysis, allowing for proactive adjustments to marketing campaigns, release schedules, and production strategies based on emerging patterns.

**Conclusion**

Our in-depth analysis of the film industry has uncovered key insights into the determinants of movie success. From the dynamic landscape of ratings evolution to the influential role of actors and directors, our findings provide actionable recommendations for industry decision-makers. Strategic release timing, genre-tailored marketing, continuous talent investment, prudent budget allocation, and adaptive strategies are pivotal for sustained success. In addition, our examination of movie length and audience preference highlights the importance of optimizing runtime based on storytelling needs, ensuring an immersive viewer experience aligned with narrative demands. By heeding these insights, stakeholders can navigate the ever-changing industry landscape, contribute to impactful cinematic experiences, and establish a lasting influence in the competitive world of film production.

**References**

Pandas Documentation (2023). <https://pandas.pydata.org/pandas-docs/stable/index.html>. Accessed on November 27, 2023.

Python Documentation (2023). <https://www.python.org/doc/>. Accessed on November 27, 2023. Accessed on November 29, 2023.

Seaborn Documentation (2023). Seaborn: Statistical data visualization. https://seaborn.pydata.org/

Statistical Function Documentation (2023). Retrieved from <https://docs.scipy.org/doc/scipy/reference/stats.html>. Accessed on on November 27, 2023.