***REPORT******ON******IMDB Movie Analysis***

**INFERENCES**

**Strong Positive Correlation:**

There is a strong positive correlation between num\_voted\_users and num\_critic\_for\_reviews (0.63), indicating that movies with more critic reviews tend to have more user votes.

num\_voted\_users also shows strong positive correlations with gross (0.65) and movie\_facebook\_likes (0.54), suggesting that movies with higher gross earnings tend to have more user votes and more likes on Facebook.

**Moderate Positive Correlations:**

There is a moderate positive correlation between movie,\_facebook\_likes and num\_critic\_for\_reviews (0.68),implying that movies with more critic reviews tend to have more Facebook likes.

duration has a moderate positive correlation with num\_voted\_users (0.31) and imdb\_score (0.26), suggesting that longer movies might attract more user votes and have a slightly higher IMDb score.

**Weak Correlations**:

Several attributes like budget, director\_facebook\_likes, and aspect\_ratio exhibit weak correlations with most other attributes in the dataset.

**Negative Correlation**:

There is a negative correlation between title\_year and aspect\_ratio (-.37), indicating that over the years, the aspect ratio might have changed.

**Miscellaneous Insights**:

actor\_1\_facebook\_likes and cast\_total\_facebook\_likes have a very high correlation (0.95),indicating a strong relationship between the total likes of the cast and the likes of the first actor.

facenumber\_in\_poster shows negligible correlations with other attributes,

indicating that the number of faces on the movie poster doesnot strongly affect other variables.

**Data Analytics Tasks**

**A. Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.**

**Task:** Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

1. **Common Genres:** The most frequent genres in the dataset are Comedy, Action, and Drama, with Adventure, Crime, Biography, and Horror also being relatively common.
2. **Genre IMDB Score Analysis:**
   * **Biography** movies tend to have the highest average IMDB scores, followed by **Documentary** and **Crime** movies.
   * **Horror** and **Thriller** films show lower average scores compared to other genres.
   * **Comedy** films exhibit a moderate average IMDB score, slightly below the overall dataset average.
3. **Variability in Scores:**
   * Genres like **Family**, **Film-Noir**, **Game-Show**, **Music**, **History**, and **Romance** have limited variability in their scores (indicated by NaN values for variance).
   * **Musical**, **Sci-Fi**, **Thriller**, and **Western** genres have notable variability in their IMDB scores (indicated by higher standard deviation and variance values).
4. **Range of Scores:**
   * **Documentary** and **Film-Noir** genres have the narrowest score ranges (0.0), suggesting consistency in ratings.
   * **Family** and **Horror** genres show a broad range of scores, indicating significant diversity in audience perceptions.
5. **Mode of Scores:**
   * The mode, representing the most frequently occurring score, varies across genres. For instance, **Film-Noir** and **Game-Show** genres have a mode of 7.6, while **Comedy** has a mode of 6.4.

These inferences provide a glimpse into how different genres tend to fare in terms of IMDB scores and the variability within each genre's ratings.

**B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.**

**Task:** Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

The descriptive statistics for movie durations provide valuable insights into the distribution of film lengths:

1. **Count:** There are a total of 5043 movies in the dataset.
2. **Mean:** The average duration of movies is approximately 106.88 minutes.
3. **Standard Deviation:** Movie durations have a standard deviation of approximately 25.83 minutes, signifying the extent of variability from the mean.
4. **Minimum:** The shortest movie in the dataset has a duration of 0 minutes, which could be an outlier or missing data.
5. **25th Percentile (Q1):** 25% of the movies have a duration less than or equal to 93 minutes.
6. **Median (50th Percentile or Q2):** 50% of the movies have a duration less than or equal to 103 minutes. This is also the measure of central tendency that divides the data into two equal halves.
7. **75th Percentile (Q3):** 75% of the movies have a duration less than or equal to 118 minutes.
8. **Maximum:** The longest movie in the dataset has a duration of 511 minutes.

In this dataset, the variance indicates relatively moderate variability in movie durations.

Identify the relationship between movie duration and IMDB score.

Movies around 100 minutes(80 to 120) duration has a high imdb score.

**C. Language Analysis: Situation: Examine the distribution of movies based on their language.**

**Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

1. **English**: The majority of movies in the dataset are in English, accounting for a significant portion with 4704 movies.
2. **Other Common Languages**: French, Spanish, Hindi, Mandarin, and German languages also have a substantial number of movies, although much fewer than English.
3. **Less Common Languages**: Some languages have relatively few movies in the dataset, such as Panjabi, Urdu, Vietnamese, Slovenian, and others, each represented by only one movie.

This information gives an overview of the distribution of movies by language in the dataset.

Analyze impact on the IMDB score using descriptive statistics.

* **English**: With 4704 movies, the average IMDb score is approximately 6.40, with a standard deviation of around 1.12. The minimum IMDb score is 1.6, and the maximum is 9.5.
* **French**: There are 73 French movies in the dataset, with an average IMDb score of approximately 7.04. The scores range between 4.9 and 8.4, with a standard deviation of roughly 0.73.
* **Spanish**: The dataset contains 40 Spanish movies, having an average IMDb score around 6.94, ranging from 4.4 to 8.2, with a standard deviation of about 0.86.

**Interpretation:**

* Languages like Portuguese, German, Hebrew, and Dutch tend to have higher average IMDb scores (above 7.3) compared to English.
* Languages such as Tamil, Russian, and Hindi have a wider range of IMDb scores, suggesting more variation in the quality of movies within those languages.
* Some languages have limited data (like Zulu, Dari, Kazakh, etc.), making it challenging to draw conclusive insights due to the small sample size.

This information provides an overview of IMDb scores across different languages, which could hint at potential trends or preferences in movie quality across various linguistic categories.

**D. Director Analysis: Influence of directors on movie ratings.**

**Task:** Identify the top directors based on their average IMDB score .

1. John Blanchard - Average IMDb score: 9.5
2. Sadyk Sher-Niyaz - Average IMDb score: 8.7
3. Mitchell Altieri - Average IMDb score: 8.7
4. Cary Bell - Average IMDb score: 8.7
5. Mike Mayhall - Average IMDb score: 8.6
6. Charles Chaplin - Average IMDb score: 8.6
7. Raja Menon - Average IMDb score: 8.5
8. Majid Majidi - Average IMDb score: 8.5
9. Damien Chazelle - Average IMDb score: 8.5
10. Ron Fricke - Average IMDb score: 8.5

These directors have earned high IMDb scores on average for the movies they've directed, indicating a consistency in delivering quality content.

Analyze directorr contribution to the success of movies using percentile calculations.

Top of Form

1. **Consistency in High Ratings:** All top directors have their 25th, 50th (median), and 75th percentile scores at the same rating. For instance:
   * Cary Bell, Mitchell Altieri, and Sadyk Sher-Niyaz have quartile percentile scores of 8.7 throughout.
   * Charles Chaplin, Mike Mayhall, Damien Chazelle, Majid Majidi, Raja Menon, and Ron Fricke also have quartile percentile scores of 8.6 or 8.5 consistently across all percentiles.
2. **High Median and Consistency:** The 50th percentile (median) scores for all directors are notably high and uniform, indicating a consistent level of quality across their movies.
3. **Low Variability:** The identical scores across quartile percentiles suggest low variability or consistency in the movies' IMDb ratings for these directors. This signifies that a majority of their films have garnered high IMDb ratings consistently.
4. **Exceptional Performance:** John Blanchard, with a 25th, 50th, and 75th percentile rating of 9.5, displays a remarkable consistency in achieving exceptionally high ratings across his directed movies.

In summary, these top directors consistently produce movies with high IMDb ratings, showcasing their ability to maintain quality and viewer appreciation across their filmography.

**E. Budget Analysis: Explore the relationship between movie budgets and their financial success.**

**Task:** Analyze the correlation between movie budgets and gross earnings

Correlation between movie budgets and gross earnings: 0.12350580785730707

The correlation coefficient between movie budgets and gross earnings is approximately 0.1235. This value suggests a positive correlation between the two variables, albeit a relatively weak one. However, in this case, a value of 0.1235 suggests a relatively low positive linear relationship between movie budgets and their resulting gross earnings. Therefore, while there might be a tendency for higher budget movies to earn more, the correlation isn't notably strong based solely on this correlation coefficient.

Identify the movies with the highest profit margin.

There are nearly 260 movies with 100% profit margin