**Project: Investigating TMDb for Movies Dataset:**

This project entails examining a dataset and presenting our conclusions. We'll employ Python libraries like NumPy, pandas, and Matplotlib to streamline our analysis. The dataset under scrutiny is the TMDb Movies Dataset, which comprises details on 10,000 movies sourced from The Movie Database (TMDb), encompassing user ratings and revenue statistics.

**Data processing:**

1. Begin by loading the TMDb Movies Dataset and thoroughly comprehend its contents.
2. Eliminate any duplicate entries from the dataset to ensure data integrity.
3. Discard unnecessary columns that are irrelevant to our analysis.
4. Rectify any inaccurate data types to ensure consistency and accuracy.
5. Remove rows containing null or zero values to maintain data quality and reliability.

**Dataset:**

The dataset comprises details on 10,000 movies sourced from The Movie Database (TMDb), encompassing user ratings and revenue statistics.

Some columns, such as 'cast' and 'genres', feature multiple values separated by pipe (|) characters. Additionally, the last two columns ending with "\_adj" indicate the budget and revenue of each movie adjusted for inflation up to 2010 dollars.

Variables:

* id
* imdb\_id
* popularity
* budget
* revenue
* original\_title
* cast
* homepage
* director
* tagline
* keywords
* overview
* runtime
* genres
* production\_companies
* release\_date
* vote\_count
* vote\_average
* release\_year
* budget\_adj
* revenue\_adj

**Files:**

tmdb-movies.csv

**Asked Questions:**

1. Are movie ratings improving, declining, or remaining steady over time?
2. Has the film industry been consistently profitable throughout the years?
3. How does the number of movies release correlate with the passage of time?
4. What is the average duration of movies?
5. Which genre boasts the highest quantity of films?
6. Is there a correlation between a larger production budget and increased profitability and popularity in films?

**How Investigation Done:**

After refining and purifying our dataset, we're prepared to delve into exploration. We'll utilize statistical computations and visual representations to tackle the research inquiries outlined in the Questions section. Employing methods like groupby(), mean(), and value\_counts(), we've responded to our queries and augmented our dataset with a new column for profit calculations via inset(). Subsequently, we'll visualize our inquiries using plot() and draw conclusions based on the outcomes.

**Answered Question:**

1. What genres consistently attract the highest popularity annually?
2. What attributes are commonly linked to movies that achieve significant revenue?
3. Is there a correlation between higher vote counts and better ratings for movies?
4. How do the most prevalent genres produced in 2000 compare to those in 2015?
5. How has the volume of film production evolved over the years?

**Findings:**

The initial research inquiry, "Which genres are most popular from year to year?" yielded unexpected findings, as the predominant genre varied significantly. Surprisingly, in only 11 instances did the most frequently produced genres align with those voted as the best genre by users, while in the remaining 40 instances, there was a disparity between the two values.

The second research query, "What characteristics are associated with movies generating high revenues?" revealed intriguing outcomes. Notably, the numeric attributes of "popularity, budget, and vote\_count" displayed the strongest correlations. While it may be argued that higher-budget movies tend to garner greater revenues, the correlation observed isn't notably high. Moreover, a higher vote\_count seems to correspond with higher revenue, although this may not necessarily indicate a reliable predictor of high revenue films. Notably, movies generating high revenue are frequently directed by Matt Damon, feature Tom Cruise as an actor, and fall within the action genre.

The third research question, "Did movies with higher vote counts receive better ratings?" did not demonstrate a clear relationship between vote\_count and vote\_average. Even when considering columns with over 2000 vote\_count, the correlation did not suggest that higher vote counts lead to higher vote averages.

Regarding the fourth research query, "What were the most popular produced genres in 2000 compared to 2015?" the data indicates that in 2015, dramas were the most frequently produced genre, followed by thrillers and action films. Conversely, in 2000, thrillers led as the most produced genre, followed by action and comedies. A visual representation demonstrates a notable increase in the number of movies produced in 2015 compared to 2000.

The fifth research question, "How did the volume of film production change over time?" revealed a significant increase in the number of films produced from 1960 to 2015. Prior to 1983, the annual production did not exceed 100 movies, whereas in 2014, 700 movies were produced. A substantial rise in movie production is observed between 1997 and 2009. It's important to note that the line chart may not provide precise results, as numerous rows from the original dataset were eliminated due to missing data.

All findings are constrained by the dataset's limitations, and as advanced statistical analyses were not conducted, these results should be considered as indicators rather than generalizable conclusions. Additionally, it's crucial to acknowledge that many entries in the dataset were removed due to missing data.

**Top of Form**

**Resources:**

**Bottom of Form**