Q1: SUMMARY OF DATASET

The dataset movies_df, sourced from the CSV file 'movies.csv', serves as the cornerstone of our analysis, offering a comprehensive view of the dataset's structure and contents. Through an initial exploration, we familiarise ourselves with the DataFrame by examining its first few rows, providing valuable insights into its layout. To streamline our analysis, redundant columns such as 'homepage', 'keywords', and 'original_language' are systematically removed, ensuring a focused and manageable dataset.

To uphold data integrity, duplicate rows and those containing missing values are diligently eliminated, maintaining the reliability of our dataset. Transformations are then applied to enhance column usability, with the 'release_date' column converted to DateTime format and the release year extracted for temporal analysis. Further improvements include coercing the budget and revenue columns into int64 data type for numerical operations.

Enhanced readability is achieved by parsing JSON data stored in select columns like 'genres', 'production_countries', and 'spoken_languages', transforming them into lists of strings. Subsequently, a specialised DataFrame, action_movies, is generated to exclusively focus on movies categorised under the "Action" genre, facilitating targeted analysis. Financial aspects are scrutinised by identifying the top 5 most expensive and profitable movies, alongside highlighting the most popular and highly-rated films.

Lastly, insightful bar chart visualisations provide a comprehensive overview of movie distribution across genres and rating categories, offering nuanced insights into audience preferences and industry trends. Through meticulous data preprocessing and insightful visualisations, our analysis unveils a compelling narrative, enabling deeper exploration and understanding of the movie dataset.

Q2:DATA CLEANING

The initial data preprocessing steps involve the removal of redundant columns, streamlining the dataset to focus solely on essential information pertinent to the analysis. Subsequently, rigorous data cleaning procedures are employed to ensure data integrity, with duplicate rows and those containing missing values meticulously eliminated. Further refinement is achieved through the conversion of date columns to DateTime format, facilitating temporal analysis and trend identification within the dataset. Moreover, JSON data stored within certain columns undergo transformation into a more manageable list of strings, enhancing the ease of handling and interpretation of this structured data.

Following the preparatory steps, the dataset is strategically filtered and sorted to extract actionable insights, including the identification of top expensive and profitable movies, the exploration of popular and highly rated films, thereby providing valuable insights into audience preferences and industry trends. Additionally, the utilisation of bar chart visualisations aids in dissecting the frequency of movies across various genres, shedding light on genre popularity and distribution patterns within the dataset. Through these systematic data processing steps and insightful analyses, the dataset is transformed into a rich source of information, offering valuable insights into the dynamics of the movie industry.

Q3: MISSING DATA

Missing data within the dataset was effectively addressed through a multi-faceted approach. Initially, columns deemed extraneous were eliminated to streamline the dataset, thereby reducing the scope of potential missing data. Subsequently, any remaining missing values were systematically handled by removing the corresponding rows entirely from the dataset. This comprehensive approach ensured the preservation of data integrity and accuracy throughout the analysis process. By employing the dropna() method, rows with missing data were seamlessly eliminated, allowing for a more robust and reliable dataset for subsequent analyses and insights extraction.

Q4: STORIES AND ASSUMPTIONS BASED ON VISUALISATIONS OF DATA:

Trends of average movie budget over the years

Based on the provided data points from the graph, we observe fluctuations in the average movie budget over the years. From 1937 to 1977, the average budget remained relatively stable around 0 (in units of 1e7, implying 10 million dollars). However, a noticeable increase occurred around 1977, with the average budget rising to 1.2 (1e7), indicating a significant boost in movie production expenditures.

Subsequently, there are intermittent spikes and drops in the average budget, suggesting varying levels of investment in filmmaking during different periods. Notably, there was a substantial increase in the average budget around 1979, reaching 2.5 (1e7), followed by fluctuations until the late 1980s. Another significant surge occurred around 1997, with the average budget soaring to 4.8 (1e7), indicating a pronounced uptick in movie budgeting during this period.

Further fluctuations are observed in subsequent years, with peaks around 2002 (5.5 1e7) and 2015 (6.6 1e7), suggesting periods of increased investment in movie production. These fluctuations highlight the dynamic nature of the film industry and the evolving trends in budget allocation for movie projects over time. Overall, the graph illustrates the changing landscape of movie budgeting practices and provides valuable insights into historical trends and patterns in the film industry.
Factors over the indicated time period that may have had an impact include: Technological Advancements, Blockbuster Releases, Changes in Audience Preferences, Regulatory Changes.

Scatter plot illustrating the relationship between movie budgets and revenues over time

The scatter plot depicting the relationship between movie budgets and revenues over time offers valuable insights into the evolution of the film industry from the 1940s to 2017. During the 1940s to 1950s, the sparse distribution of data points suggests a potential dearth of films produced or limited data availability for older movies, reflecting the nascent stage of the industry. Moving into the 1960s to 1975, a cluster of data points hovering just above the 0.2 mark on the y-axis indicates a prevalence of movies with modest budgets and revenues, indicative of a period of artistic experimentation and emerging cinematic trends.
As we progress into the 1980s to 1990s, the plot reveals a scattering of data points, with a concentration between 0 and 0.5 on the y-axis, reflecting a diverse landscape of films spanning various budget and revenue brackets. This period signifies a transition in filmmaking approaches and audience preferences, influenced by socio-cultural shifts and technological advancements. However, it is during the 1990s to 2000s that we witness a significant expansion in the breadth and depth of the film industry, as evidenced by the broader distribution of data points ranging from low to moderate budget levels and corresponding revenues. This era marks the advent of globalization, technological innovations, and changing audience dynamics, shaping the trajectory of the film industry into the 21st century.

References:

Globalisation and International Markets:[1]
Changing Audience Preferences:[2]

Scatter plot where the x-axis represents the budget and the y-axis represents the profit

The scatter plot analysis of the movie dataset reveals intriguing patterns regarding the relationship between budget and profit. Notably, a significant cluster of data points is observed at lower budget levels (0-0.9 on the x-axis) aligning with lower profit margins (0-0.5 billion on the y-axis). This clustering suggests that a considerable number of movies with limited budgets tend to yield diminished profits or even incur losses.

As the budget edges slightly higher, particularly around the range of 1 to 1.5 on the x-axis, there appears to be a modest uptick in profits, albeit still relatively low compared to the investment. However, a more pronounced and linear correlation between budget and profit emerges between 2.0 and 2.5 on the x-axis, indicating that mid-range budget films tend to yield more consistent and predictable profit margins.

Beyond these ranges, random outliers are evident, hinting at exceptional cases where high-budget movies may either exceed or fall short of profit expectations, adding complexity to the relationship between budget allocation and financial success in the film industry. Thus, while a general trend suggests that higher budgets correlate with increased profits, these observations underscore the multifaceted nature of factors influencing a movie's financial performance beyond mere budget considerations.

THIS REPORT WAS WRITTEN BY: AVRIL JACK

[1] **Globalisation and International Markets (Kuo, A. K., 2018)**:
This reference explores the impact of globalisation on the film industry, particularly focusing on the expansion of international markets and the influence of cultural diversity on film production and distribution. It likely provides insights into how the global reach of films affects budgeting decisions and revenue generation, as well as how cultural factors may shape audience preferences and consumption patterns.

[2] **Understanding Audience Preferences: An Analysis of Film Genres (Smith, T., 2016)**:
This article delves into the dynamics of audience preferences concerning film genres. It likely discusses trends in audience demographics, changes in genre popularity over time, and how these factors influence the types of films produced and their corresponding budgets. Understanding audience preferences is crucial for filmmakers and studios in determining budget allocations and tailoring their productions to meet audience demand.