

**TASK**

**Exploratory Data Analysis on the Movies Data Set**

[](https://www.hyperiondev.com/)

**Introduction**

**Summary of the data set**

The movies dataset encompasses diverse films, varying in budget, genres, and production aspects. It includes information on original language, popularity, production companies, and countries. The dataset also features details like release dates, revenues, runtimes, and vote averages. With insightful metrics such as vote count, it provides a comprehensive view of the film landscape, offering valuable insights into the cinematic world's financial, creative, and audience engagement dimensions.

**DATA CLEANING**

**Summary of the methods and visualisations done during data cleaning**

* **Columns Removal:**
  + Identified and removed unnecessary columns using **drop(columns=columns\_to\_remove)**.
  + Columns removed: 'homepage', 'keywords', 'original\_language', 'original\_title', 'overview', 'production\_companies', 'status', 'tagline'.
* **Duplicate Removal:**
  + Removed duplicate rows using **drop\_duplicates()** to ensure data integrity.
* **Handling Missing Values:**
  + Checked for missing values using **isnull().sum()** to identify the count of missing values in each column.
  + The summary of missing values was likely used to make decisions on how to handle them. Methods might include imputation or removal of rows/columns with missing values.
* **Visualisation**:
  + Visualizations include a bar chart revealing genre distribution, a scatter plot illustrating the relationship between budget and revenue, and another bar chart displaying the average revenue for each genre in the "movies.csv" dataset.

**MISSING DATA**

The code snippet **print(movies\_df.isnull().sum())** is employed to identify and quantify missing values in the "movies.csv" dataset. The output provides a comprehensive view of columns with missing data, crucial for subsequent handling strategies, such as imputation or removal.

**DATA STORIES AND VISUALISATIONS**

* **Genre Distribution:**
  + A bar chart was created to visualize the distribution of movie genres in the dataset. This provides insights into the popularity of different genres.
* **Budget vs. Revenue:**
  + A scatter plot was generated to explore the relationship between budget and revenue. This helps understand the financial performance of movies.
* **Average Revenue by Genre:**
  + Calculated and visualized the average revenue for each genre. This sheds light on the financial success of different movie genres.

**THIS REPORT WAS WRITTEN BY : Mbali Matches**

