

**TASK**

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

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

**Introduction**

This is a data set of movies that contains information about the profitability of each movie, we will perform data clean and exploratory analysis to better understand the data and visualize which movie that has more profit and least profit through graphs attached on the project.

**DATA CLEANING**

While performing data cleaning, we observed that the data set contains duplicates values, we removed the duplicates with same values by performing such query e.g df = df[df[‘budget’] > 0].

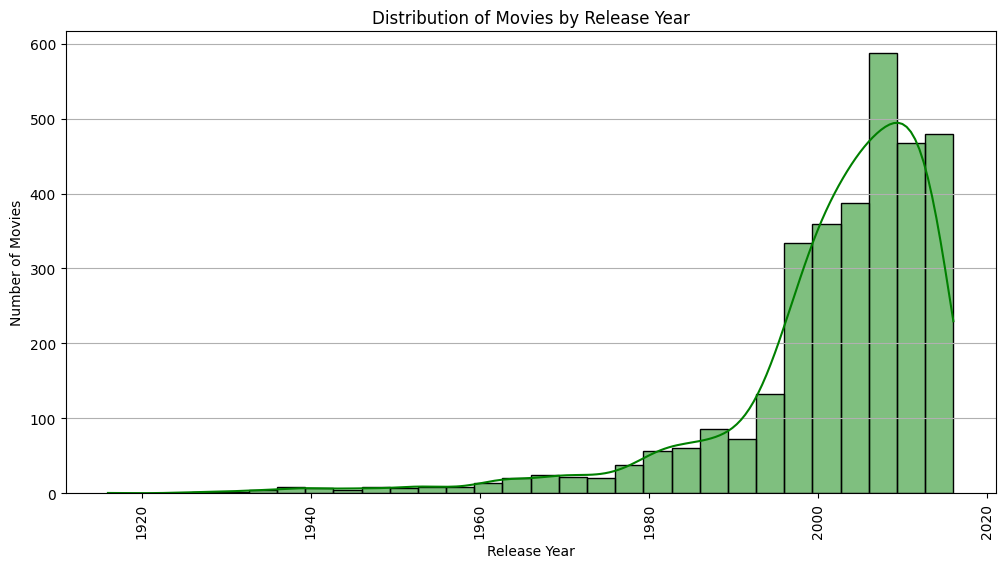
**MISSING DATA**

We observed that the data set has missing values so we performed the query in as follows e.g df = df [(df [‘budget’] > 0)] & (df [‘revenue’] > 0]).

**DATA STORIES AND VISUALISATIONS**

Data Story:

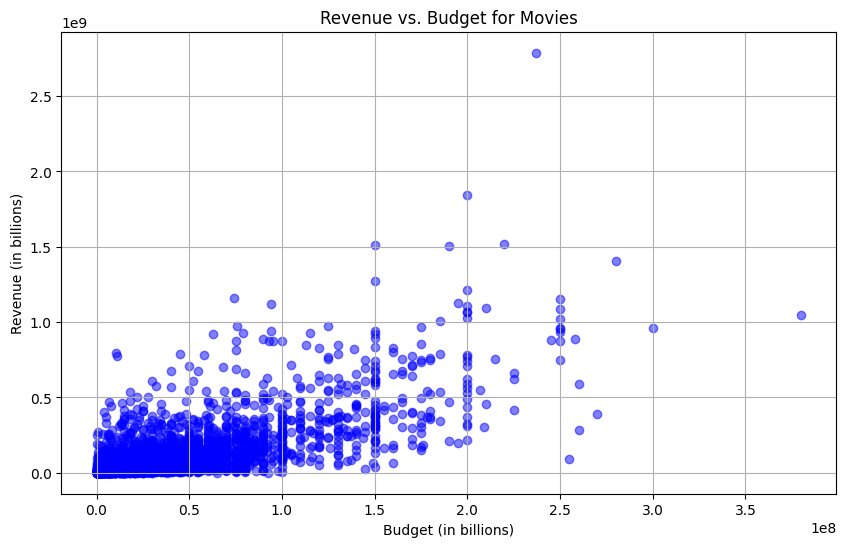
The greater number of movies distributed has generally increased over the years, with the most movies released in recent years. The early 2000s saw a significant drop in number of movies been distributed, but it has since recovered as it gained popularity from 2000s upwards.

****

**Data Story:**

There is a positive correlation between movie budget and revenue, indicating that movies with higher budgets tend to generate higher revenues. However, the correlation is not perfect, suggesting that factors other than budget also influence revenue.

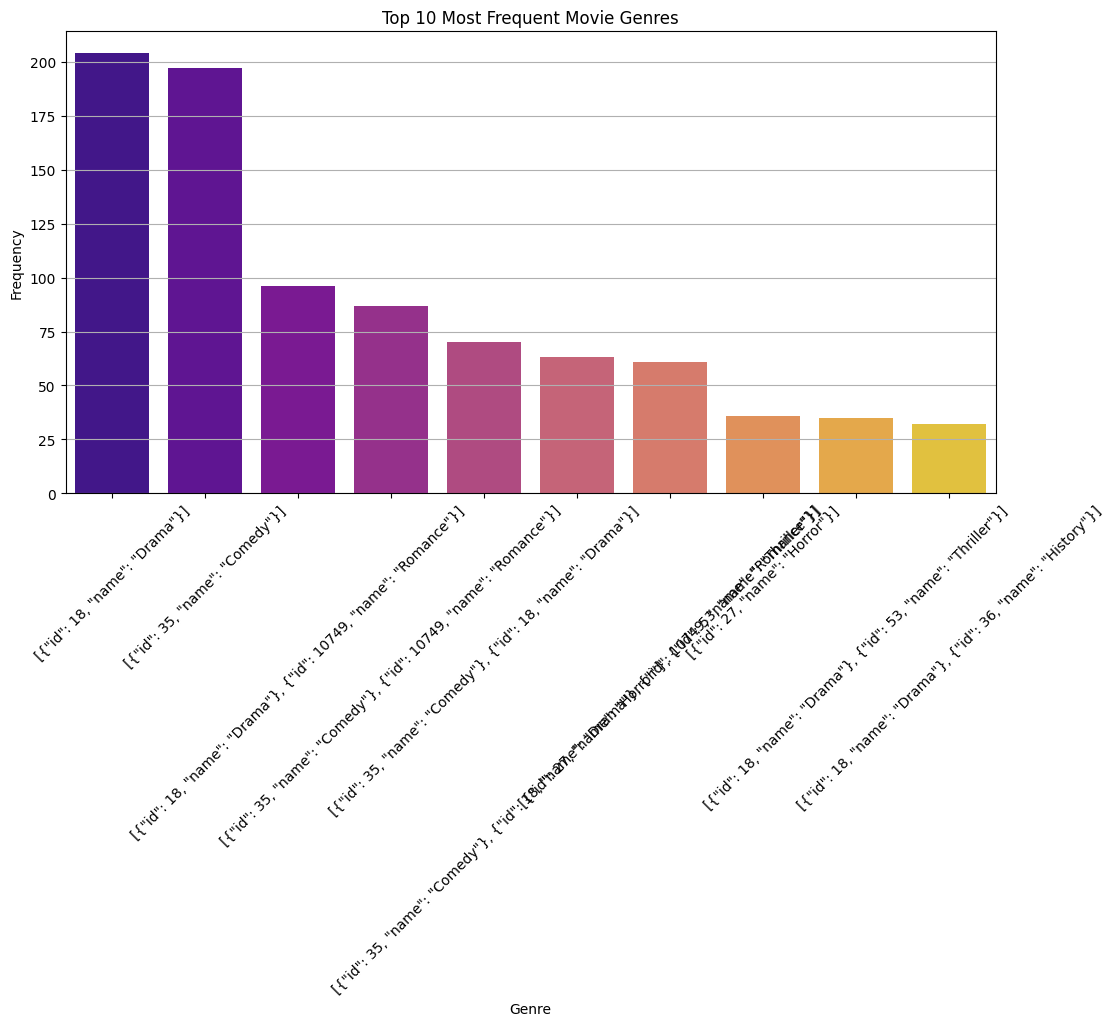
Visualization: A scatter plot with a regression line showing the relationship between budget and revenue.

****

**Data Story:**

The most popular genres in the dataset are Action, Comedy, and Drama. While Action movies tend to have the highest total profit, genres like Horror and Mystery have higher median profits.

Visualization: A bar chart showing the distribution of movies across different genres.

****

**THIS REPORT WAS WRITTEN BY : BONGANI ZULU**

