

Exploratory Data Analysis: Movie Industry Trends

INTRODUCTION

This dataset contains information on a wide range of movies, including budget, revenue, popularity, and average viewer ratings. The objective of this analysis is to identify patterns and relationships that influence a movie's financial performance and audience reception.

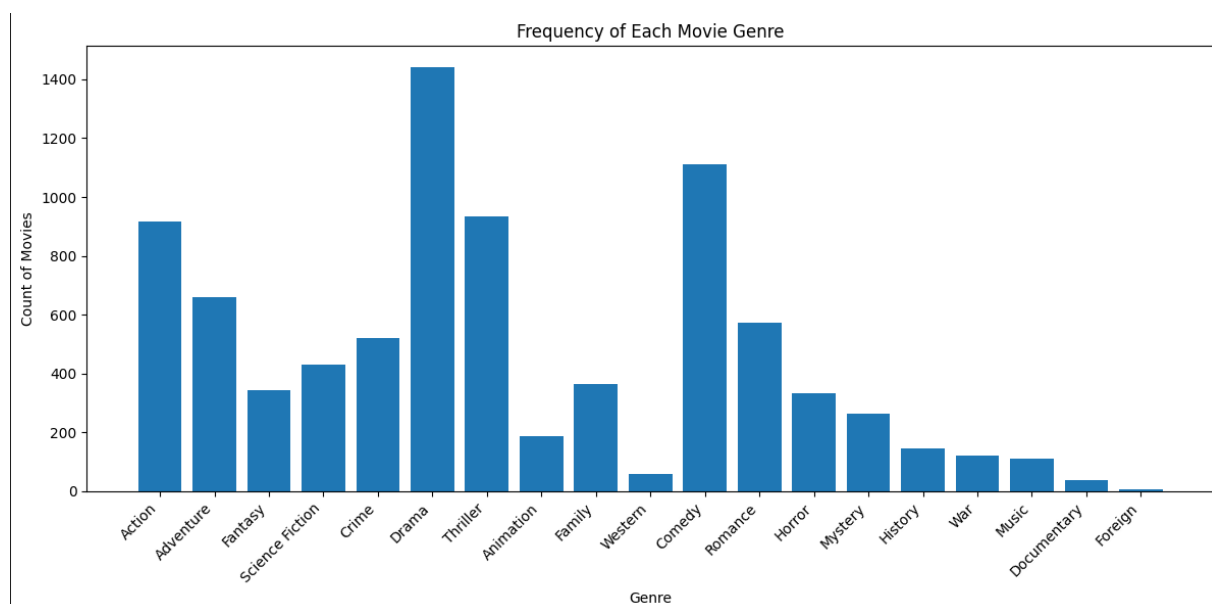
DATA CLEANING

Redundant or non-essential columns were removed, including: ['homepage', 'keywords', 'original_language', 'original_title', 'overview', 'production_companies', 'status', 'tagline']. Duplicate records were also removed to ensure data integrity.

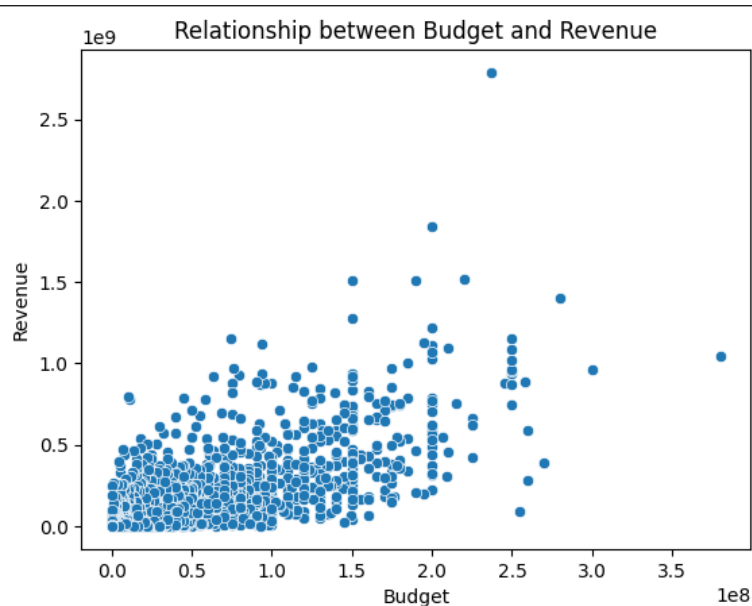
MISSING DATA

Rows with a budget or revenue value of 0 were treated as missing data, as these values indicate unrecorded information. These rows were removed from the analysis.

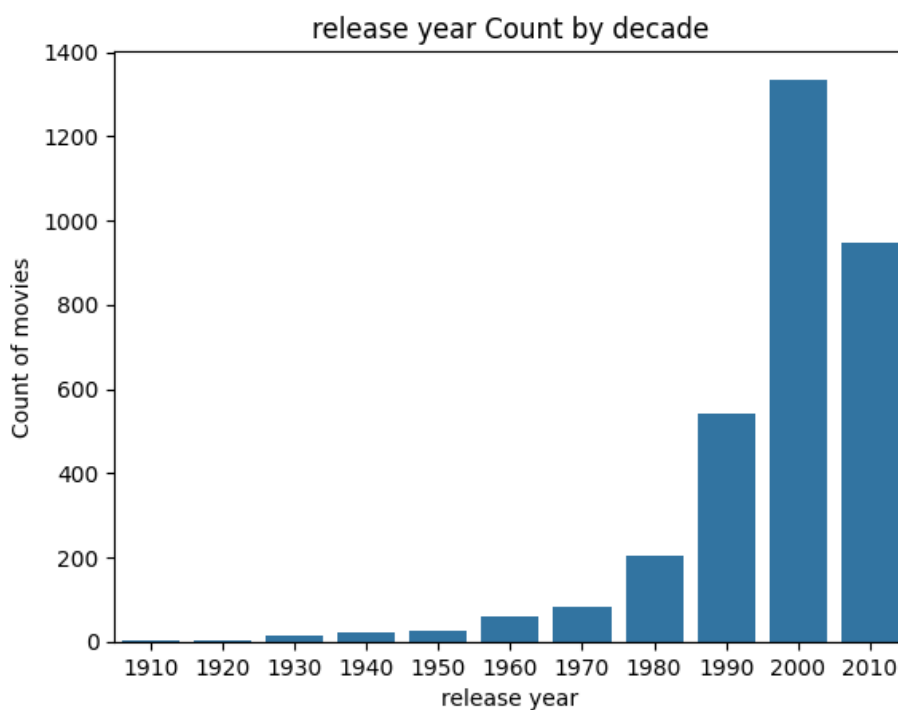
DATA STORIES AND VISUALISATIONS



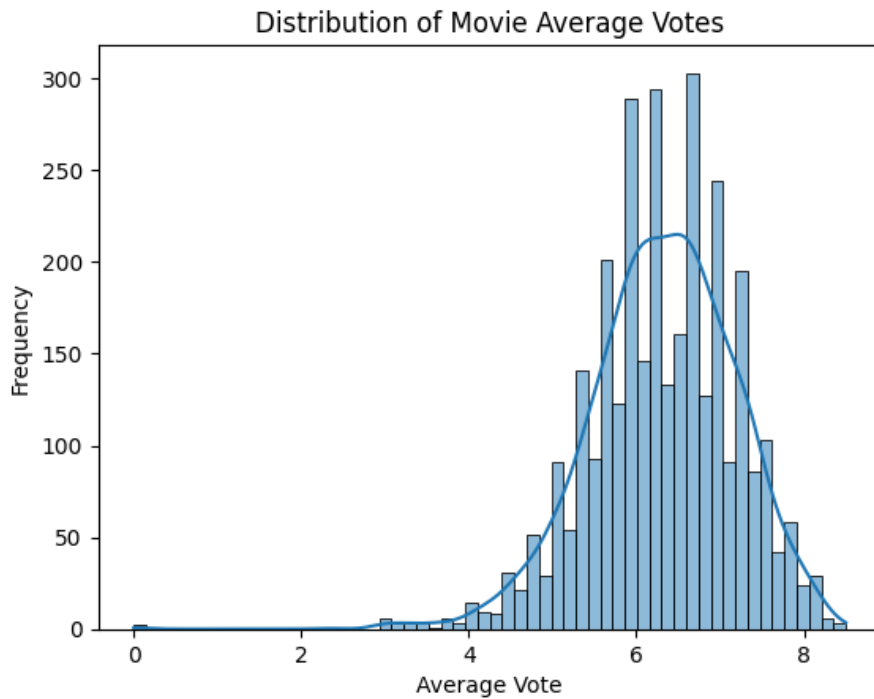
This bar plot shows the distribution of movie genres in the dataset. Drama, Action, Thriller, and Comedy dominate, indicating that these genres are the most commonly produced and widely consumed.



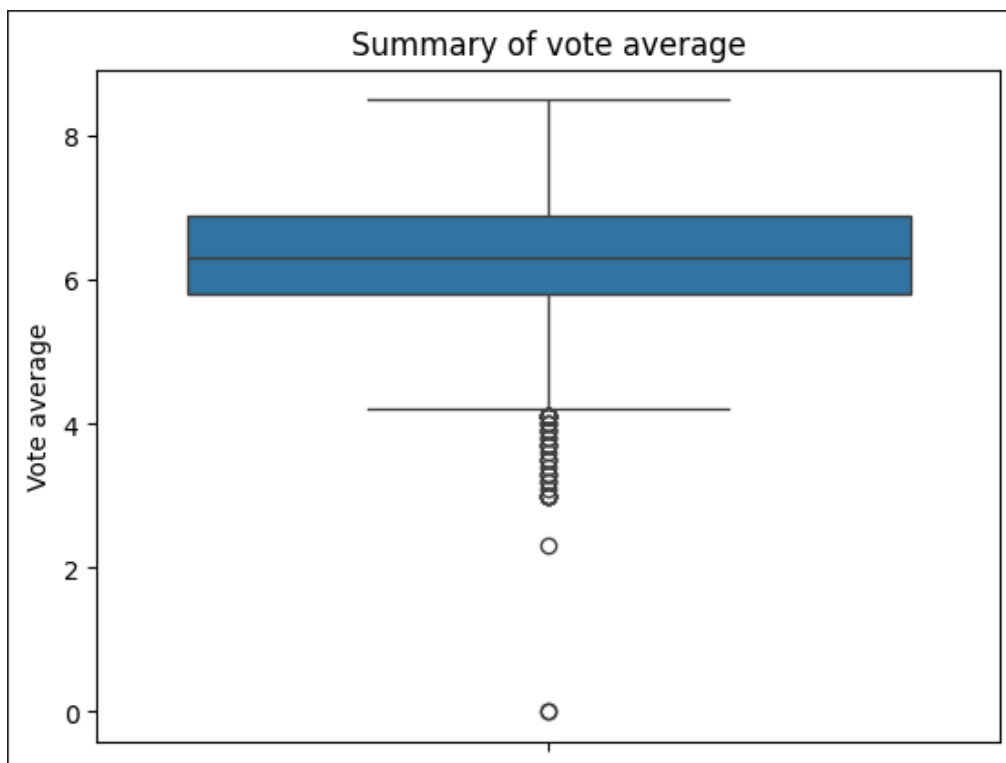
This scatter plot reveals a positive correlation between a movie's budget and its revenue, suggesting that higher production budgets generally increase the likelihood of higher earnings.



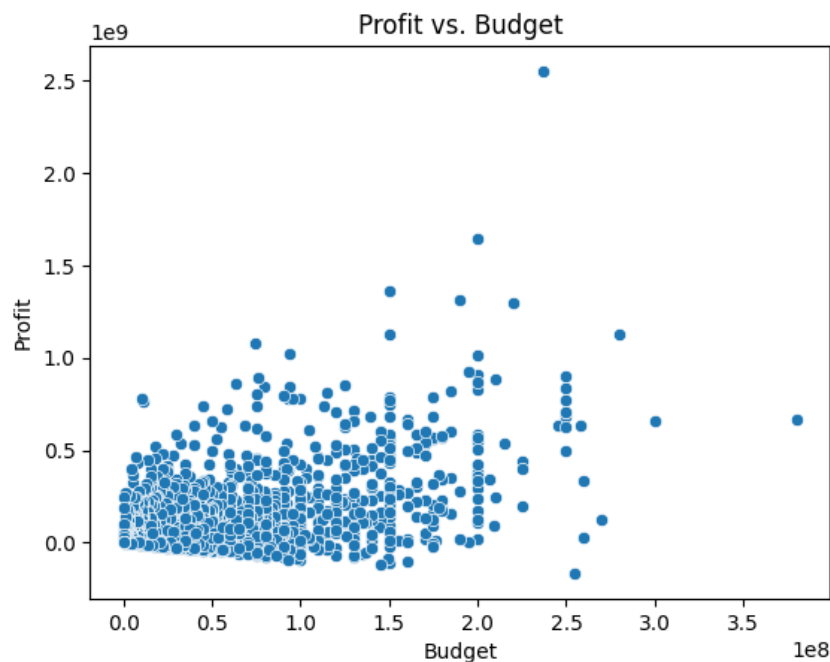
This visualization highlights a significant increase in movie production over time, with output more than doubling from the 1980s to the 1990s and increasing further into the 2000s.



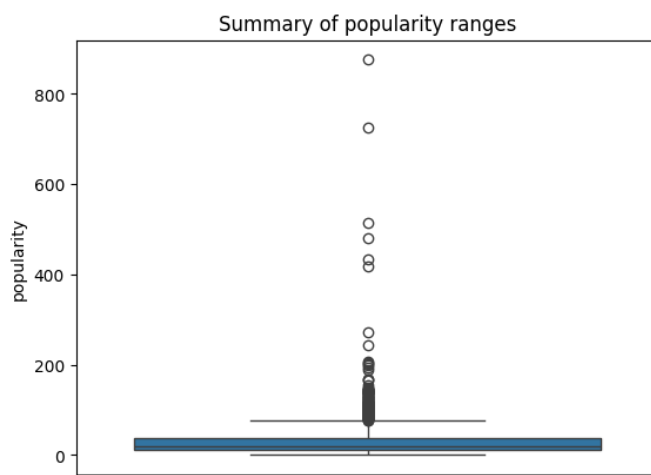
Most movies have ratings between 4 and 8, forming a roughly normal distribution, which suggests that extreme ratings are less common.



The majority of vote averages fall between 6 and 7, with several lower-rated outliers extending down to ratings around 2.



This plot shows a positive relationship between budget and profit, indicating that higher-budget movies tend to generate higher profits, although variability increases with budget size.



Most movies have low popularity scores, but there are notable outliers with extremely high popularity, reaching values close to 900.

Prepared by: Christiaan Cronje | Data Analytics Portfolio