

**TASK**

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

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

**Introduction**

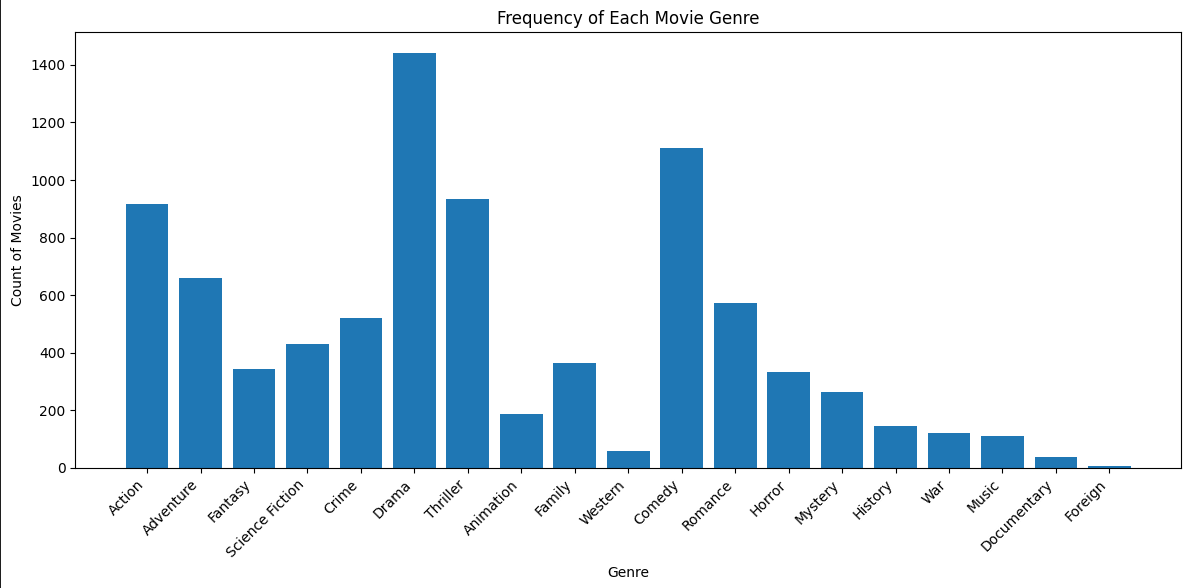
This dataset consists of numerous movies with their details such as budget, revenue, popularity, vote average etc. The aim of this analysis is to find the patterns within this dataset to know the factors that influence each movies performance.

**DATA CLEANING**

I removed redundant or unnecessary columns that won’t be needed such as the following: ['homepage', 'keywords', 'original\_language', 'original\_title', 'overview', 'production\_companies', 'status', 'tagline']. All duplicate rows were removed.

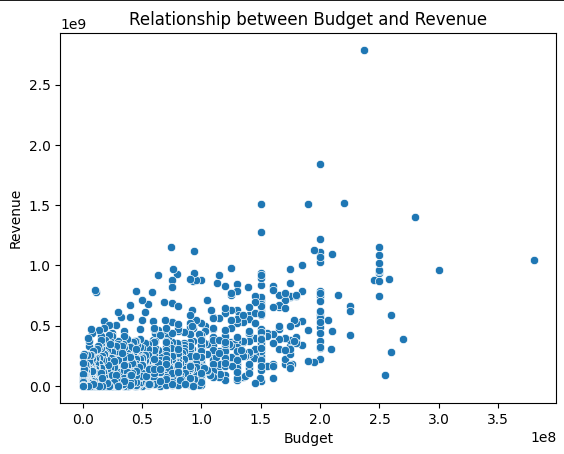
**MISSING DATA**

Rows that show 0 budget or revenue imply that the values have not been recorded. These rows were discarded.

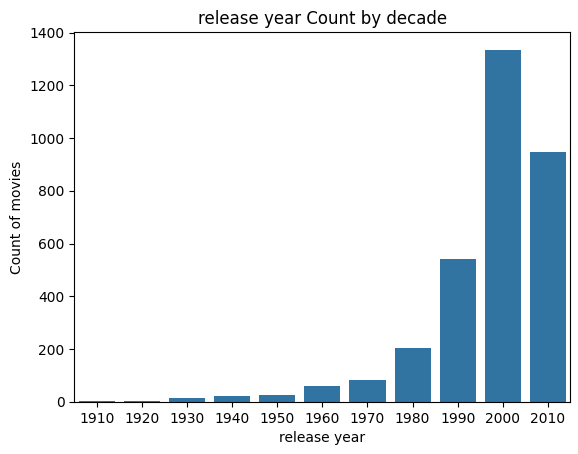
**DATA STORIES AND VISUALISATIONS**

We can see that most movies fall under the following categories: Drama, Action, Thriller and Comedy. These are all genres that most people are familiar with.

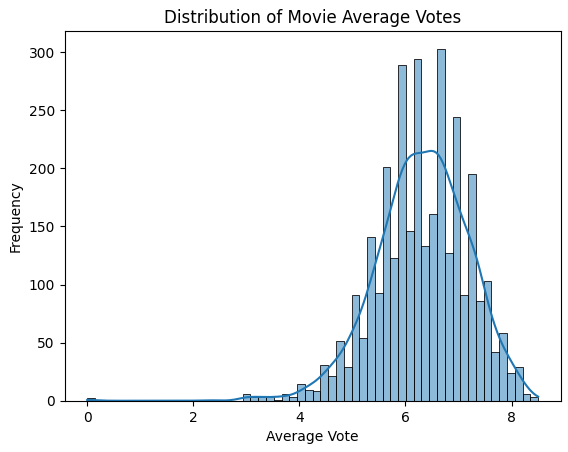
It is safe to say that there is a positive correlation between a movie’s revenue and budget, meaning that the more budget a movie has, the more likely it is to generate more revenue.



From the 1980’s to the 1990’s the production of movies increased by over 100 percent and from the 1990’s to the 2000’s it increased by a further 100 percent meaning these last few decades the count of movies drastically skyrocketed.



It seems that most movie ratings fall between the range of 4 to 8 out of 10, being normally distributed.



The majority of movie voted average in the dataset fall within the ranges of roughly 7 to 6, but there are a lot more low average outliers, going down to around a rating of 2.

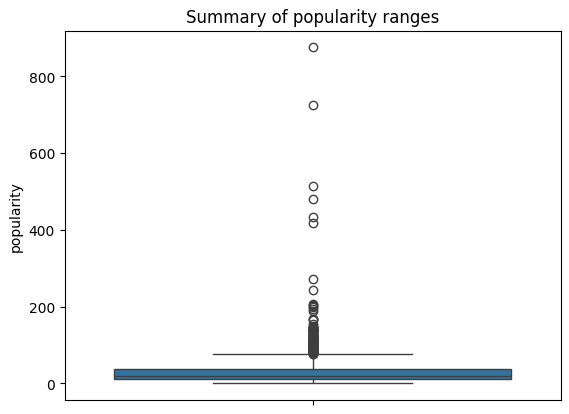
**A diagram of a voting graph

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

We can see that there is a positive correlation between a movie’s profit and budget, meaning that the more budget a movie has, the more likely it is to generate more profit.

****

Most movies in the dataset have a very low popularity score, being between roughly 10 and 40. However, there are a significant number of outliers with movies that are considerably more popular than the average movie in this dataset, with some of them reaching up to nearly 900.

**THIS REPORT WAS WRITTEN BY : Christiaan Cronje**

