

**TASK**

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

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

**Introduction**

**This dataset contains data about several popular movies. The dataset contains information about each movie such as the genre, revenue, production companies and so on.**

**DATA CLEANING**

**SUMMARY OF THE METHODS AND VISUALISATIONS DONE DURING DATA CLEANING**

When loading the dataset and viewing it, several columns needed to be removed to make the analysis of the data a lot easier. These columns contained unnecessary data that made the understanding of the data complicated. Following that all rows with duplicate data were removed and rows that had missing data were removed. A few of the columns data types were changed to be able to perform our EDA. The ‘release date’ column was changed to Datetime, and a new column was added called ‘release year’ which contained the release year of the respective movies. The ‘budget’ and ‘revenue’ columns data type were changed using numpy’s int64 method. Columns ‘genres’, ‘production countries’ and ‘spoken language’ were in JSON format so were flattened into a format that could be easily read.

**MISSING DATA**

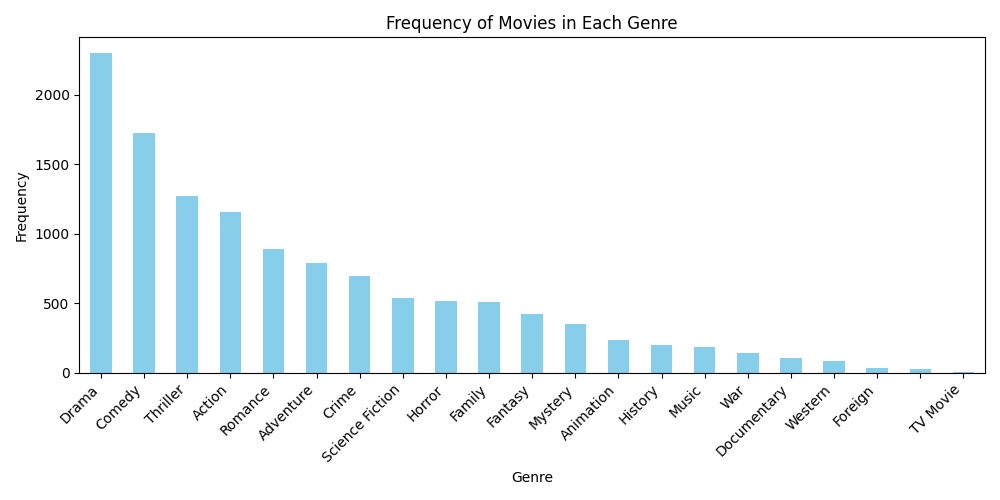
**ANY MISSING DATA? HOW DID YOU HANDLE IT?**

Any rows that had missing data were removed as mentioned above.

**DATA STORIES AND VISUALISATIONS**

**THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALISATIONS OF THE DATA**

When creating a visualisation of the frequency of movies in each genre, it was interesting to see that drama was ranked as the highest rated out of all the genres.



**Budget vs Revenue Scatterplot:**

When analysing the scatterplot, it’s interesting to see that the greater the budget of a movie, then the greater the revenue of the movies. This could be down to many factors such as a greater budget means the studio is able to hire bigger "A list" actors and actresses which draws larger audiences to their movie. Another reason could be that the bigger the budget then the better the production of the movie can be in regard to special effects and so on which will no doubt boost the revenue.

A graph showing a budget vs revenue

Description automatically generated with medium confidence

**Genre Popularity Bar Plot:**

One of the interesting things noted in this bar plot is Adventure ranks at the top of the average popularity. When considering so my high-quality action movies that have been produced over the years, I would have assumed that Action would feature the highest. In terms of education, it is also worrying that Documentaries feature so low in the average popularity.

A graph of a number of people

Description automatically generated

**Number Of Movies Released Per Year Line Chart:**

An interesting observation is the number of movies seem to have drastically increased in the late 1980's to early 1990's. This could be because of the improvement of technology in the filming industry around that time. Around 2009 is where the most movies were produced and noticeably radically decreased. he global financial crisis that began in 2008 had a significant impact on various industries, including the film industry. Reduced funding, tighter budgets, and uncertainty about the economic future could have led to a decrease in movie production.

A green line graph with numbers

Description automatically generated

**THIS REPORT WAS WRITTEN BY : JASON LUKE EDWARD**

