

**TASK**

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

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

**Introduction**

The movie dataset provides a comprehensive overview of the film industry, encompassing various genres, financial successes, popularity and more. Through the exploratory data analysis (EDA), significant insights have been revealed. A positive correlation was found between movie budgets and revenue. It was also discovered that, while nominal budgets exhibit an upward trend over the years, adjusting for inflation shows the average budgets to remain relatively stable. It is further seen that genres like drama, comedy and thriller dominate industry preferences. Data pre-processing, including handling missing values and extracting relevant information, ensured the dataset’s reliability.

**DATA CLEANING**

* The following columns were deemed unnecessary for this analysis and were dropped from the DataFrame: ‘homepage’, ‘keywords’, ‘original\_language’, ‘original\_title’,’ ‘overview’, ‘production\_companies’, ‘status’ and ‘tagline’.
* Duplicate rows were removed.
* The ‘release\_date’ column was converted into DateTime format and the year was extracted and added to a separate column.
* The budget and revenue columns were converted to integer data types (int64).
* The ‘genres’, ‘production\_companies’ and ‘spoken\_languages’ columns were in JSON format and were flattened through the ‘parse\_col\_json’ function to flatten the data into a more easily interpreted format.

**MISSING DATA**

* Rows with any missing (NaN) data were removed.
* There were some movies in the database that had a budget and/or revenue of zero. It is assumed that these values were recorded incorrectly. Thus rows with a budget or revenue of zero were removed.
* Some other budgets and revenues were entered incorrectly. Some entries have missing zeroes. The missing zeroes were, however, not consistent. Some entries were missing six zeroes, whereas others were only missing one zero. Therefore, when studying the cheapest movie budgets, the 20 ‘cheapest’ entries were ignored, after which the entries appeared to be more accurate.

**DATA STORIES AND VISUALISATIONS**

Data Exploration:

* All movies of a certain genre were found (particularly action movies)
* The most and least expensive movies (based on budget) were compared.
  + A histogram of the movie budgets was plotted to see the outliers and trend of the budgets as seen in Figure 1.
* Each movie profit was calculated and the most profitable movies were examined.
* The most popular movies were examined.
* Movies rated above 7 were identified.
* The success of each genre was studied.
  + A bar plot shown in Figure 2 of the frequency of movies in each genre was generated. The most frequently occurring genre was drama, followed by comedy, followed by thriller. The least occurring genre was foreign, followed by documentary, followed by western.

A graph of a movie budget

Description automatically generated

Figure 1

A graph of different colored bars

Description automatically generated

Figure 2

Visualisations and data stories:

1. Relationship between Revenue vs Budget:

* Figure 3 shows a scatterplot of the revenue versus the budget of each movie in the log scale. From this we are able to identify a generally positive correlation between the two variables, particularly for higher budgets and revenues where the correlation is stronger. This shows that higher budget movies tend to generate higher revenue.

A graph with blue dots

Description automatically generated

Figure 3

1. Movie release trends over the years:

* A line chart, shown in Figure 4, of the number of movies released each year was plotted. This provided insights into the growth or decline in the film industry over time. From this we were able to see the slow increase in the release of movies from ~1920 to ~1990. During this time the movie industry was still in its early stages of development. After 1990 the number of movies released increases dramatically till ~2010 this could have been due to the rise of blockbuster films, greater international collaboration, and new distribution channels such as home video and cable television. In ~2010 the rate of movie releases reaches a peak where it begins to rapidly fall again. This could have been a result of competition from streaming services and an increased popularity in television series as opposed to movies.

A graph showing the number of movies released

Description automatically generated

Figure 4

1. Budget vs Year: Nominal budget vs Inflation Adjusted Budget:

* A scatter plot was made of the movie budgets (in the log scale) vs release years. The budgets were then adjusted to account for inflation and the results were compared. The nominal budget appeared to have a positive trend with time. Interestingly, however, when examining the inflation adjusted budgets, there is not much increase in the average budgets. There is increased budget variability in more recent years, likely due to advancements in technology and creative opportunities. However, the average budget is relatively stable.

A chart of a graph

Description automatically generated with medium confidence

Figure 5

**THIS REPORT WAS WRITTEN BY : Kathleen Sellick**

