

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

The dataset contains information on 4,803 movies.

The budget column has a mean value of approximately $29 million, with a standard deviation of around $40 million. The minimum budget is $0, and the maximum budget is $380 million.

The popularity column has a mean value of 21.49, with a standard deviation of 31.82. The minimum popularity score is 0, and the maximum popularity score is 875.58.

The revenue column has a mean value of approximately $82 million, with a standard deviation of around $163 million. The minimum revenue is $0, and the maximum revenue is $2.79 billion.

The runtime column has a mean value of 106.88 minutes, with a standard deviation of 22.61 minutes. The minimum runtime is 0 minutes, and the maximum runtime is 338 minutes.

The vote\_average column has a mean value of 6.09, with a standard deviation of 1.19. The minimum vote average is 0, and the maximum vote average is 10.

The vote\_count column has a mean value of approximately 690, with a standard deviation of around 1234. The minimum vote count is 0, and the maximum vote count is 13,752.

The dataset also includes other columns such as genres, production\_countries, release\_date, spoken\_languages, title, and id, which provide additional information about each movie.

**DATA CLEANING**

The following methods and visualizations were performed during the data cleaning process:

1. Columns removal: Identified and removed redundant or unnecessary columns from the dataset using the drop() method. The columns removed were ['keywords', 'homepage', 'status', 'tagline', 'original\_language', 'overview', 'production\_companies', 'original\_title'].

2. Duplicate rows removal: Checked for duplicate rows in the dataset using the duplicated() method and removed them using the drop\_duplicates() method. The number of duplicate rows was also calculated.

3. Zero budget and revenue removal: Discarded rows where the budget or revenue was equal to zero using boolean filtering. The number of rows with zero budget or revenue was calculated.

4. Date manipulation: Converted the 'release\_date' column to the DateTime format using pd.to\_datetime(). Extracted the release year from each release date and created a new column 'release\_year'.

5. Format conversion: Converted the 'budget' and 'revenue' columns to integer format using the astype() method with np.int64 as the argument.

6. Flattening JSON columns: Flattened the 'genres', 'production\_countries', and 'spoken\_languages' columns, which were in JSON format, into a format that can be easily interpreted. This was done using lambda functions and list comprehensions to extract the desired values.

These methods and visualizations were performed to clean and transform the dataset, making it more suitable for analysis and interpretation.

**MISSING DATA**

Based on the dataset, it appears that the dataset contains movies with zero budget or zero revenue, indicating missing or unrecorded values. To handle this missing data, the code filters the dataframe movies\_df and keeps only the rows where both the budget and revenue are not equal to zero.

The steps taken to handle the missing data are as follows:

1. The initial number of rows in the dataset is obtained using the shape attribute of the dataframe (num\_rows1).

2. The code filters the dataframe movies\_df using the condition (movies\_df['budget'] != 0) & (movies\_df['revenue'] != 0). This keeps only the rows where both the budget and revenue values are not equal to zero, effectively discarding entries with missing or unrecorded values.

3. The resulting dataframe after filtering is assigned back to movies\_df.

4. The number of rows in the filtered dataframe is obtained using the len() function (num\_rows2).

5. The difference between the initial number of rows and the filtered number of rows gives the number of entries with zero budget or revenue (zero\_budget\_revenue).

Therefore, the missing data represented by zero budget or zero revenue values is handled by removing the corresponding entries from the dataframe.

**DATA STORIES AND VISUALISATIONS**

To extract stories and assumptions based on the visualizations of the data, we can analyze the provided information:

1. Most Expensive Movies:

The analysis of the most expensive movies reveals that high-budget movies can generate significant revenue. The top five most expensive movies include "Pirates of the Caribbean: On Stranger Tides," "Pirates of the Caribbean: At World's End," "Avengers: Age of Ultron," "Superman Returns," and "John Carter." These movies had budgets ranging from $260 million to $380 million, with revenues ranging from $284 million to over $1 billion. This suggests that these high-budget movies can be worth the investment if they generate substantial revenue.

2. Cheapest Movies:

The comparison of the cheapest movies highlights the diverse nature of the film industry. The movies "Modern Times," "A Farewell to Arms," "Split Second," "Bran Nue Dae," and "The Prophecy" had budgets ranging from $1 to $8. Their revenues varied significantly, from a few dollars to millions. This shows that low-budget movies can also have varying levels of success, with some generating modest profits despite their minimal budgets.

3. Top 5 Most Profitable Movies:

The analysis of the top 5 most profitable movies reveals their financial success. "Avatar," "Titanic," "Jurassic World," "Furious 7," and "The Avengers" achieved remarkable profitability. These movies had budgets ranging from $150 million to $237 million and generated revenues ranging from $1.3 billion to $2.7 billion. This indicates that these movies not only recovered their budgets but also earned substantial profits, making them highly successful ventures.

4. Movies Rated Above 7:

Analyzing movies rated above 7 allows us to identify highly regarded films. The high-rated movies included in the dataset cover various genres and production countries. Some notable examples are "Avatar," "The Dark Knight Rises," "Tangled," "Avengers: Age of Ultron," and "Harry Potter and the Half-Blood Prince." These movies received favorable ratings (above 7) and garnered significant popularity, as reflected in their high vote counts and revenues.

5. Most successful genres: The bar plot indicates that Action and Adventure are the most frequent genres, with a frequency of 250 movies. The other genres have frequencies ranging between 5 and 22.

6. Box Plot of Movie Budgets by Genre: This visualization shows the distribution of movie budgets across genres. Animation and Drama have the highest budgets, while Thriller, Horror, and Crime have relatively lower budgets.

7. Scatter Plot of Movie Revenue vs. Popularity: The scatter plot reveals that most movies have a popularity score between 0 and 200 and a revenue score between 0 and 1. Interestingly, there is no direct correlation between popularity and revenue. The most popular movie (with a popularity of 840) has a revenue of 1.5, while the movie with the highest revenue (3) has a popularity around 200.

8. Bar Plot of Movie Release Years: The graph illustrates the number of movies released over the years. The number of movies steadily increases from 1992 (with 20 movies) to a peak in 2012 (with 165 movies). However, the number of movies starts to decline after 2012, with around 70 movies released in 2016.

To summarize the key findings from the visualizations: the exploration of the data provides insights into the relationships between variables such as budget, revenue, profitability, and ratings. It showcases the range of financial outcomes in the film industry and highlights the success of high-budget blockbusters as well as the potential profitability of low-budget movies. Additionally, it underscores the popularity and critical acclaim of movies rated above 7, indicating their appeal to audiences. These visualizations provide insights into the frequency of genres, budget distribution, the relationship between revenue and popularity, and the trend in movie releases over time.

**THIS REPORT WAS WRITTEN BY : ARLINDO ALMADA**

