Exploratory Data Analysis on the Movies Data Set

Introduction

The movies.csv file was used to create the EDA report for this Capstone Project. The data is an overview of different movies released and their specifics such as their budget, genres, and other variables of interest. This was performed using the relevant modules in Jupyter.

Data Cleaning

Firstly, the relevant modules were imported, data set was loaded and DataFrame was created. In order to gain meaningful insights, the dataset had to be prepared by removing redundant and duplicate columns using the drop() and drop\_duplicates() functions respectively. Columns were reformatted; namely the date, budget and revenue columns into the relevant objects (integers) and the release\_date column into datetime object. in order to be able to manipulate the data. Lastly the columns were flattened as they were previously in JSON format.

Missing Data

There was missing data as the budget and revenue columns had some entries which were zero indicating that those values had not been recorded. Those entries were thus discarded using the dropna() function.

Data Stories and Visualisations

The relationships between the different variables were explored. The top 5 most expensive movies in terms of budget were determined which were: Pirates of the Caribbean: On Stranger Tides, Pirates of the Caribbean: At World's End, Avengers: Age of Ultron, Superman Returns and John Carter. All these movies fall under the action and adventure genres, which corroborates with the fact that these genres are the most expensive to produce respectively as illustrated in the barplot (Figure 1). These expenses may be due to costs of special effects and talents used to make blockbusters. The most expensive movies had a rating between a range of 6.1-6.9 and the cheapest movies had a rating with a rating between 6.0-6.9; these ranges overlap showing that there is not necessarily a connection between a film's budget and its quality/acclaim, and this is reinforced by the weak positive linear correlation of 0.092 between the two variables determined by the correlation matrix (Figure 2). The 5 most profitable movies were determined, and they were as follows: Avatar, Titanic, Jurassic World, Furious 7 and The Avengers which had a vote average with a range of 6.5-7.4. Whereas the least 5 profitable movies were as follows: The Lone Ranger. The Wolfman, The Alamo, Mars Needs Moms, Dragonball Evolution which had a vote average with a range of 2.9-5.9. It is easy to assume that the first impressions of the movies by enthusiasts and professional critics which may either encourage or deter other viewers from watching films and thus influence ticket sales, but it is shown otherwise as there is a weak positive correlation of 0.2 between revenue and vote average meaning that high revenue does not necessarily mean that a movie is good. A regression line plot (Figure 3) was made in order to show how the size of the budget affects the revenue earned by a movie and it has a positive slope indicating a positive linear relationship which is reinforced by the strong positive correlation of 0.73 shown in the correlation matrix. Therefore, the higher the budget, the higher the revenue that would be generated by a film.

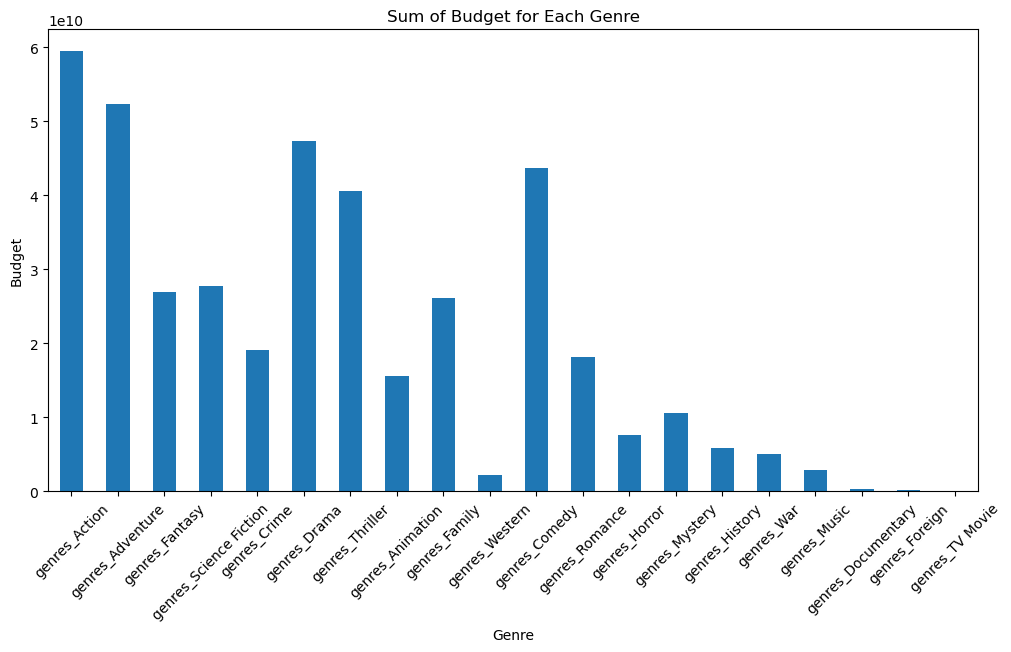


Figure 1

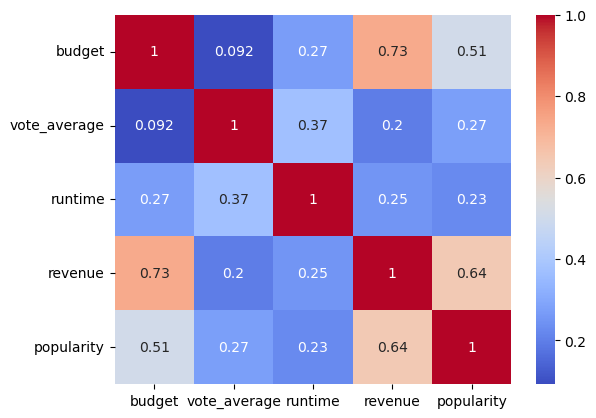


Figure 2

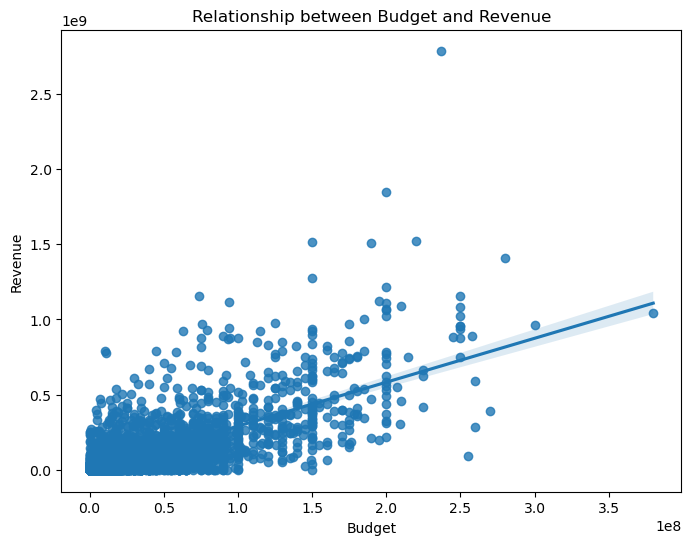


Figure 3

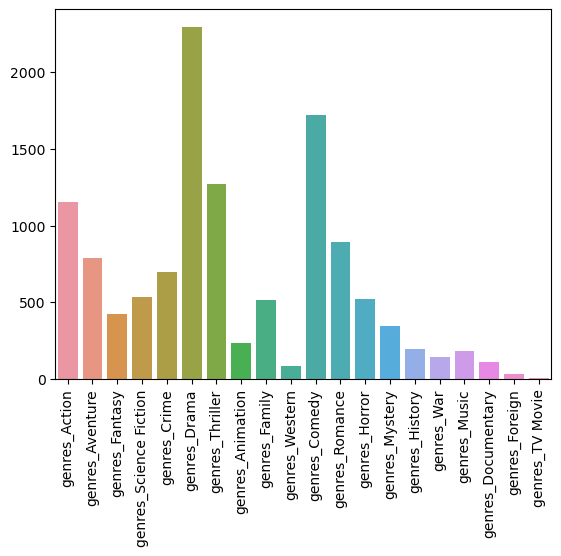
A barplot (Figure 4) was created to determine the most successful genres using seaborn module by determining the frequencies/counts of these genres. Drama was the most popular followed by comedy then thrillers. Thus, these genres are the ones that attract the most revenue, but the data shows otherwise in the barplot (Figure 5) as crime followed by western generated the most revenue. 

Figure 4

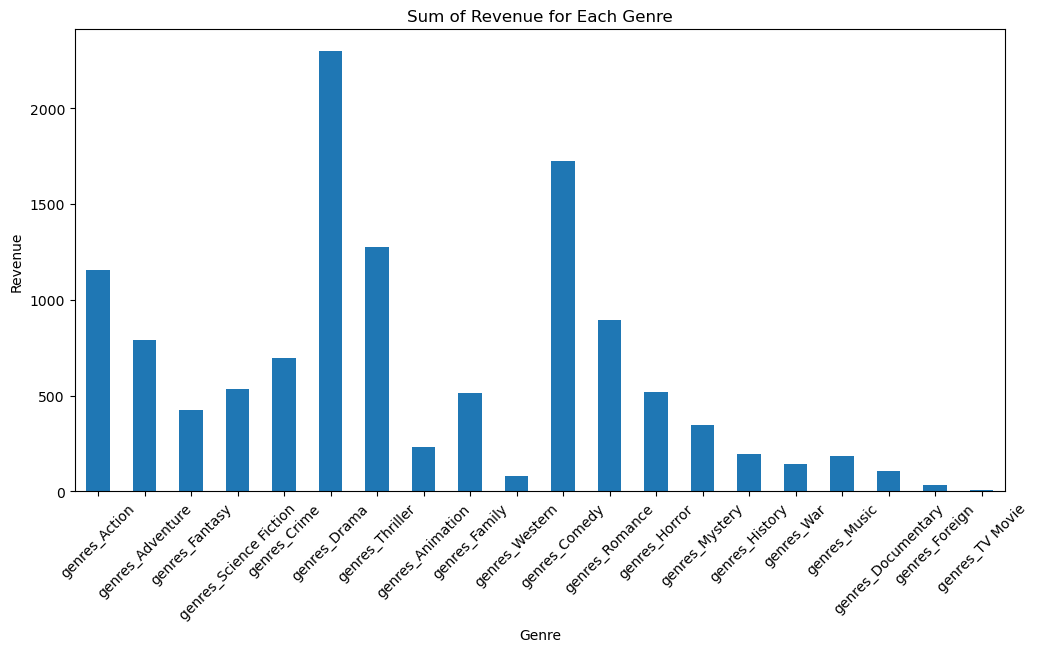


Figure 5

Other visualisations were also created. A histogram (Figure 6) was plotted of the distribution of the movie ratings. Generally, it has a bell curve shape indicating a normal distribution. Throughout the years, there has been a general downward trend in the ratings of movies as illustrated by the barplot (Figure 7), indicating a decline in quality of movies. The volume of movies greatly increased at the time of the decrease of vote averages shown, indicative of quantity over quality of movie releases. The increase of movie releases is also a factor of growth of the cinema industry caused by technological advancements, production and innovation efficiency and increased accessibility to viewers due to cultural globalisation.

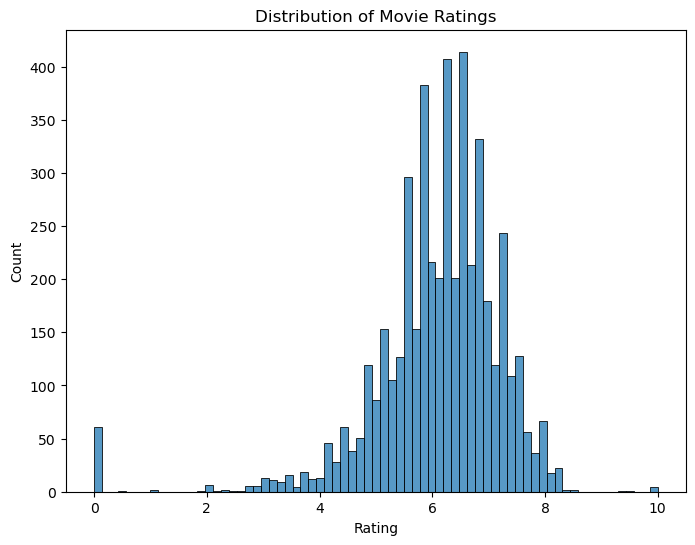


Figure 6

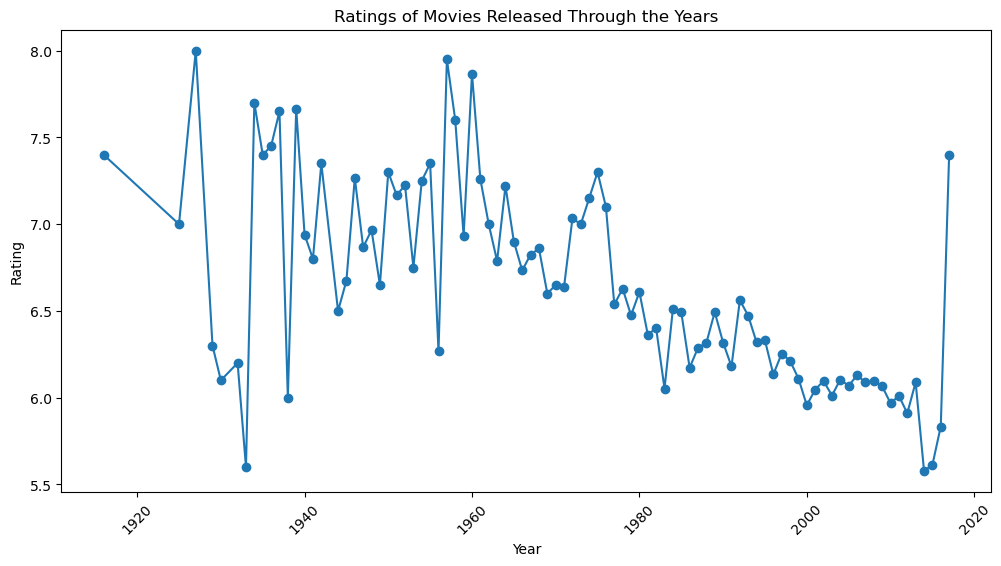


Figure 7

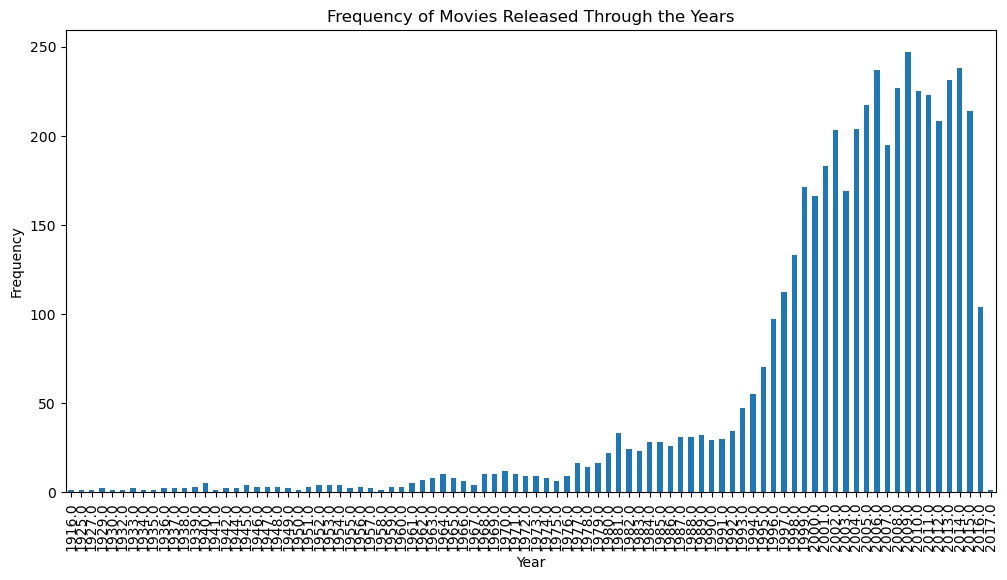


Figure 8