

**TASK**

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

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

**Introduction**

The dataset contains information about movies, the following are the titles/columns of which the information is set under: budget, genres, homepage, id, keywords, original language, original title, overview, popularity, productions companies, production countries, release date, revenue, runtime, spoken language, status, tagline, vote average and vote count. The dataset is colour coordinated which makes it easier to identify the different columns.

**DATA CLEANING**

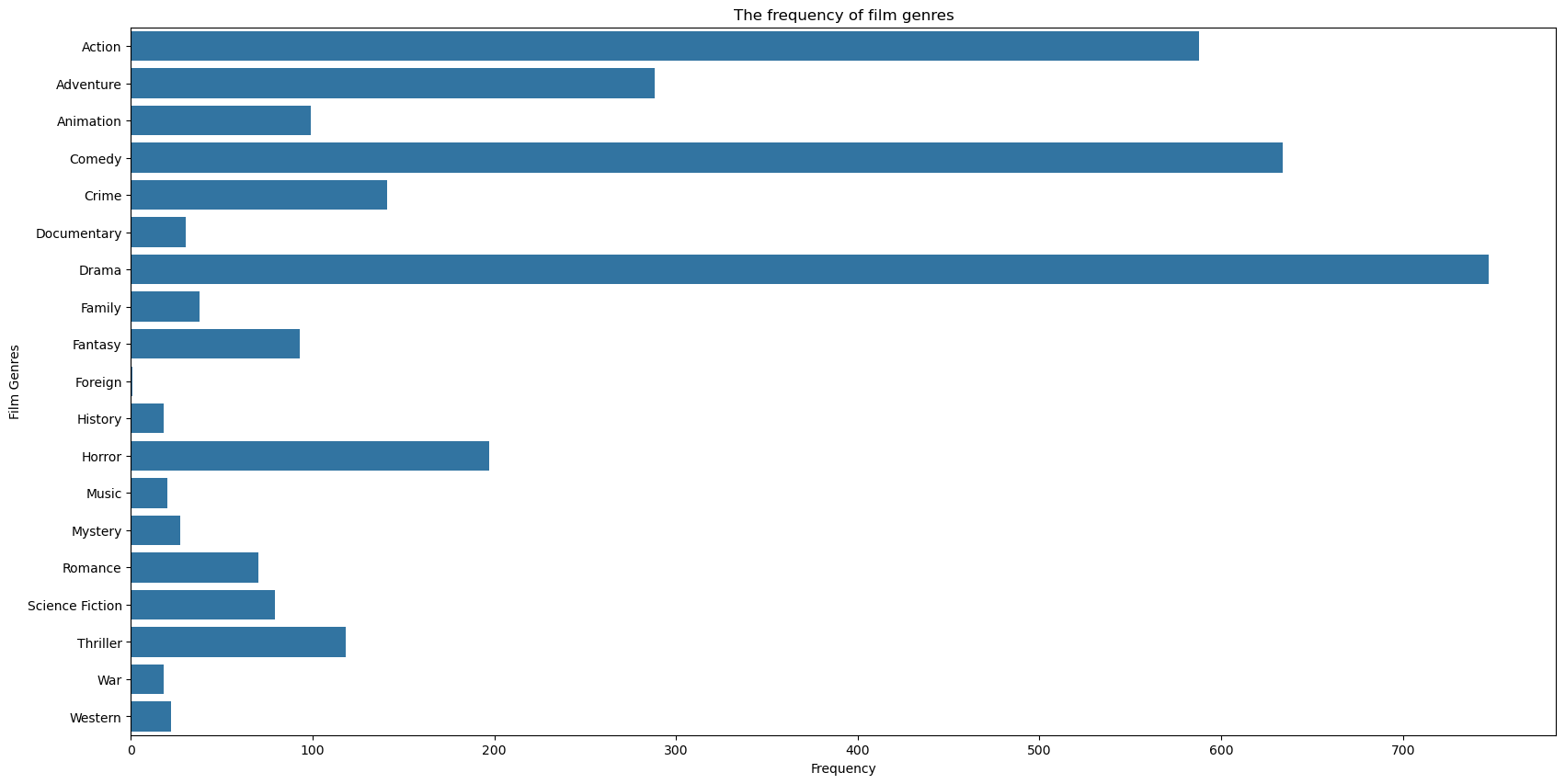
The dataset contained over 4800 data, that needed to be cleaned prior to analyses. By looking over the titles or columns, the dataset needed was identified and those that did not fall under that category was removing using that drop function. The drop function is used to remove any column/s and the following were identifies as columns to be removed: 'homepage', 'keywords', 'original language', 'original title', 'overview', 'production companies', 'status', 'tagline'. In any dataset there is a chance that there are duplicates in the rows, from the remaining columns, duplicates rows were removed using the drop duplicate function. From the dataset the budget and revenue columns were identified as the most important columns, because of this any rows that contained missing data was remove.

**MISSING DATA**

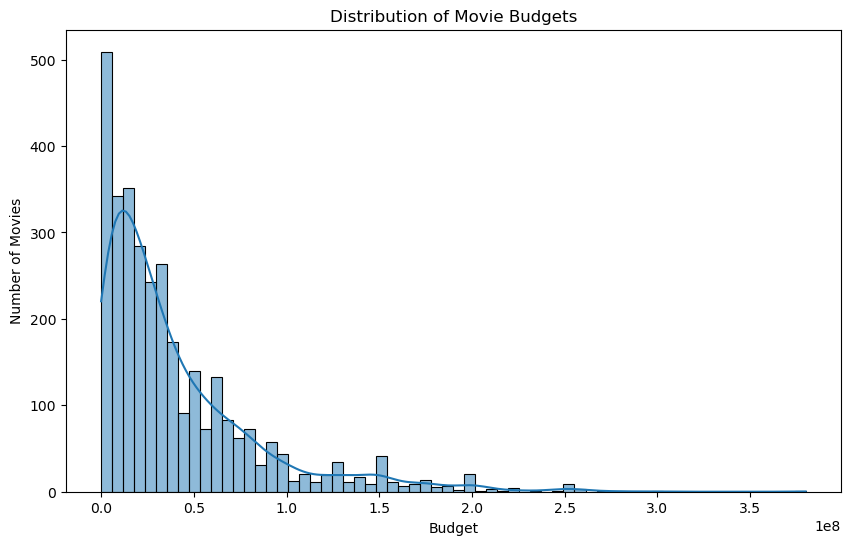
Yes, there are columns with missing data and this was dealt with in two ways. The first was to remove columns that were not in use or deemed as redundant or unnecessary were removed entirely. The second method was to remove rows with zero/ missing data from the budget and revenue columns, as these columns were important and needed to be used in analysing the data.

**DATA STORIES AND VISUALISATIONS**

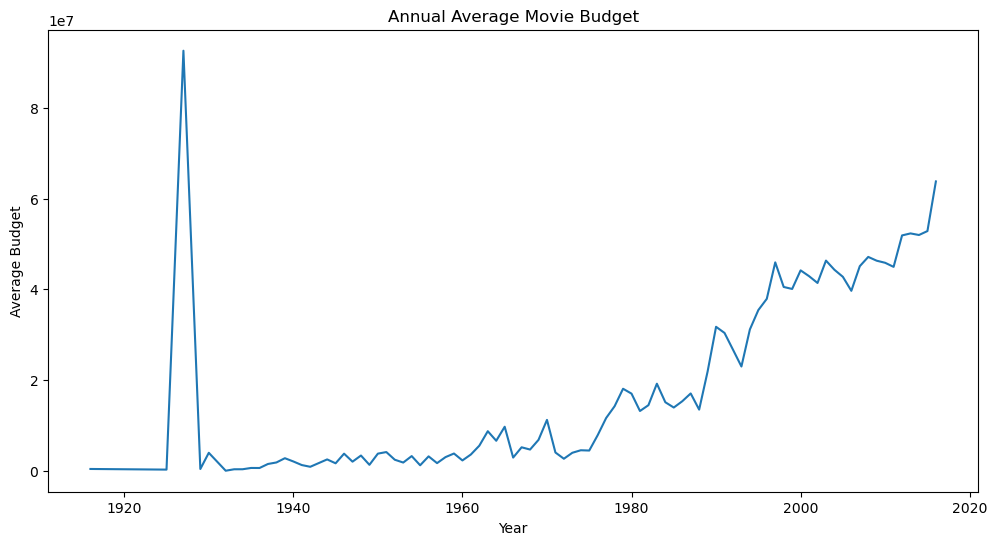
The graph below aims to show the frequency of film genres from the dataset. By using the bar-graph provides a clear visual to display the frequency of film genres. The graph shows that Drama is the most frequency genre, followed by comedy and action. It also displays that foreign, history and war are the less frequent genres in the movie database.

****

The second graph contain information about the distribution of the movie budgets. The information in the graph is displayed using a histogram for clear and easy display of the information. The graph shows that less that 50 movies get the large amounts of the budget, whereas over 400 movies have the smallest amount of budget distribution.

****

The third graph shows the annual average movie budget. The graph uses the line-graph to display the results of the annual average movie budget. The graph shows that the was a sharp increase and decrease in the average budget between 1920 and 1940 and a steady increase from 1980.

****

**THIS REPORT WAS WRITTEN BY : VANESSA TOOKA**

