

**TASK**

**Exploratory Data Analysis on the Movie Data Set**

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

This dataset contains information about the most popular films and will be used to better understand how the budget reflects on the outcome of the move

Before beginning the analysis of the dataset, it must be cleaned in order to make is more understandable. The first step in this process is to check for redundant columns which have no relevance to the dataset.

Most of the films original language was in English making the column irrelevant as we wouldn’t learn much information during analysis, which is why it has been dropped. The same applies for status as most of these movies have already been released.

Id and tagline column are all unique and they are not being used as the index for this dataset so will be dropped.

Original Title has been dropped as most of the movies have the same name, this makes the column a duplicate and irrelevant.

Overview column isn’t very concise, it contains a lot of words, making the data irrelevant.

There are several duplicate rows, these records should be dropped and is done using the Title column.

There are some movies with missing data, these are movies with 0 budget or revenue, these records have been dropped.

The format of the date column, makes the information harder to analyse as currently it is not being defined as a date, we extract the year from the release date column to analyse yearly data.

The format of the budget and revenue column also needs to be changed as analysis will be easier to conduct on integers.

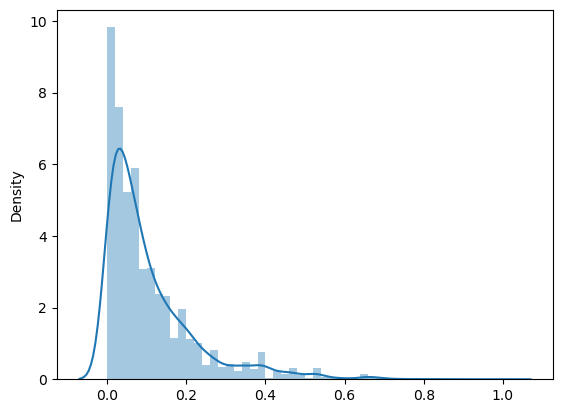
There is data in columns which is encoded in JSON, it needs to be converted into string format so it can be manipulated during analysis.

Once the data cleaning has been completed, the next stage is the exploration. This stage includes manipulating the dataframe to increase our understanding.

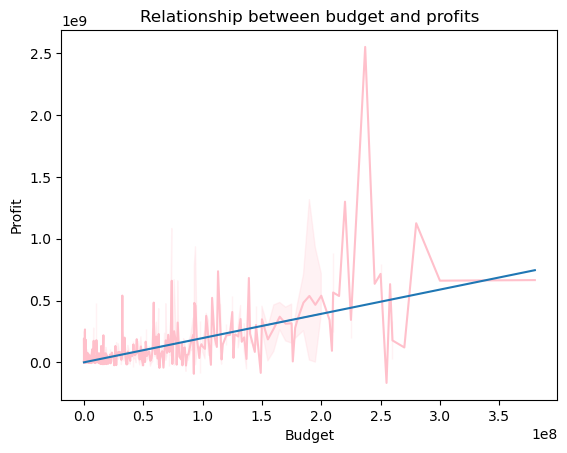
|  |  |
| --- | --- |
| BUDGET | TITLE |
| 380000000 | Pirates of the Caribbean: On Stranger Tides |
| 300000000 | Pirates of the Caribbean: At World's End |
| 280000000 | Avengers: Age of Ultron |
| 270000000 | Superman Returns |
| 260000000 | John Carter |

*The table show the movies with the highest budget.*

The difference between the most expensive movie has been scaled and displayed in the visualisation below.



We use the budget and revenue columns to calculate the profit for each of the movies, this was then used to create the visualisation below which shows the relationship between the budget and the movie. The visualisation shows there is a positive correlation between the 2 variables.



|  |  |
| --- | --- |
| PROFITS | TITLE |
| 2550965087 | Avatar |
| 1645034188 | Titanic |
| 1363528810 | Jurassic World |
| 1316249360 | Furious 7 |
| 1299557910 | The Avengers |

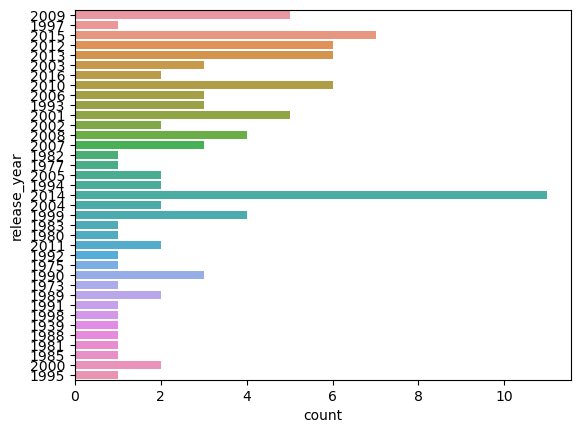
*The table shows the top 5 most profitable movies*

The difference between the most and least profitable movie is 2716675177.

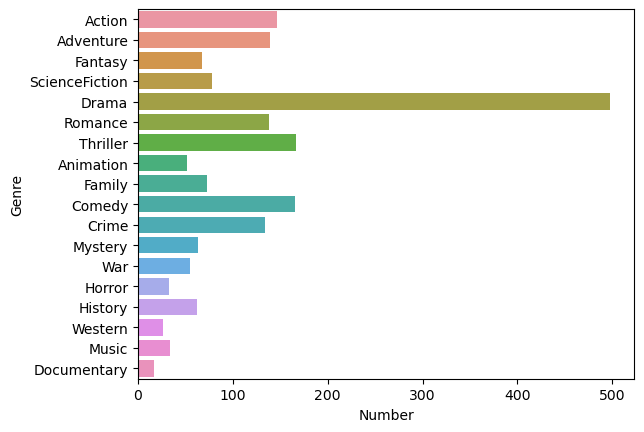
|  |  |
| --- | --- |
| POPULARITY | TITLE |
| 875.581305 | Minions |
| 724.247784 | Interstellar |
| 514.569956 | Deadpool |
| 481.098624 | Guardians of the Galaxy |
| 434.278564 | Mad Max: Fury Road |

*The table below shows the most spoken about movies.*

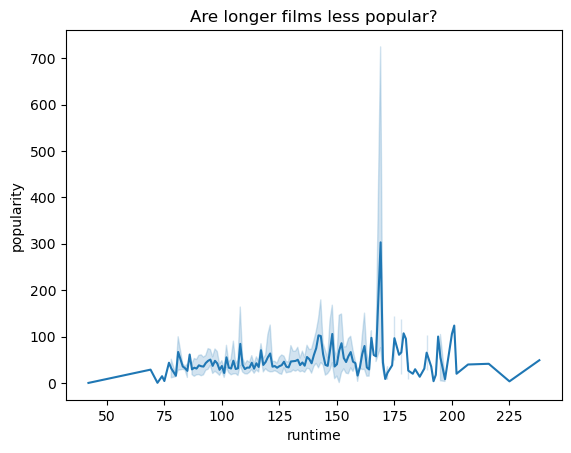
The graph below shows the years with the most profitable movies, as you can see movies before released before 2000 have the lowest profits, this could be due to the lack of technological advancement, post 2000 it became easier to access movies from home, increasing the revenue stream for the movies.



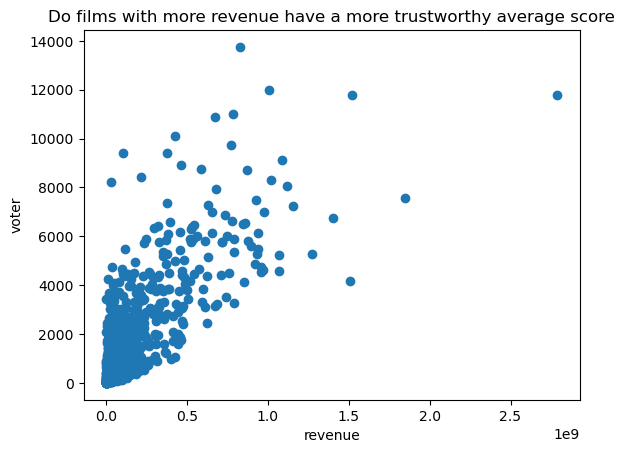
The visualisation below looks at the frequency of the genres in the dataset, we can see the most commonly created genre is Drama.

****

The next visualisation looks at the relationship between the length of the movie and it’s popularity. It shows there isn’t a relationship between the 2 variables.



The next visualisation works on the assumption that a higher revenue means more people watched the film, this would therefore mean that the average vote is more accurate. The visualisation shows a positive correlation, implying the assumption is correct.



**THIS REPORT WAS WRITTEN BY: Sara Dimtsu**

