

**TASK**

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

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

**Introduction**

Movies data set containing 20 columns and 4804 rows of information. I imported varies libraries to assist my analysis and loaded the movies.csv file into Jupyter.

**DATA CLEANING**

In summary I carried out the following:

* I identified columns that are unnecessary and removed them from the data frame
* I removed any duplicate rows, sorted and displayed results
* I changed the release date column into date format and extracted the year from the date into a new column
* I change both the budget and revenue columns into integer using numpy’s int64 method
* I changed the genres, spoken languages and production countries from JSON format to list

**MISSING DATA**

In summary I carried out the following:

* I removed any rows where the budget = 0 from the data frame
* I removed any rows where the revenue = 0 from the data frame

**DATA STORIES AND VISUALISATIONS**

In summary I carried out the following:

Budgets, Revenue and Profit

* I looked at the top 5 movies by highest and lowest budgets
* I did a comparison of budget vs revenue and found that the bigger budget the more revenue the movie will have.
* I produced a joint plot to visualise a strong correlation between the movie budget and revenue.
* I created a new Profit column and took the figures from the revenue column minus the budget column to return the value of profit per movie
* I identified the highest and lowest movies by profit
* I also identified the average profit by movie title

Movie Popularity

* I analysed the most popular movies by their popularity score
* I analysed the most voted movies by their score
* I analysed the voted average summary and created a seaborn displot to visualise.
* I analysed the movies with the highest vote average.

Movie Year

* I analysed the number of movies produced by year and notice there is an increase in in the production of movies starting in the 1990’s
* I analysed the earliest movies by year
* I analysed the year where movies made the most profit and created a plt to visualise.
* I also created a seaborn jointplot to visualise this further and compare budget, year and profit.

Movie Genres

* I analysed the genres and found Drama category had the highest movies a plt to visualise.
* I also created a seaborn barplot to visualise this further and compare genre and movies.

Movie Runtime

* I analysed the length of all the movies using displot which shows around 110 minutes on average
* I compared the length of all the movies against profit using relplot which shows a connection between runtime and profit as an upward trend occurs
* I compared the length of movies runtime by year and found that as the years have progressed movies runtimes have increased but there are still exceptions to this rule.

**THIS REPORT WAS WRITTEN BY : LOUISE RANDALL**

