

**TASK**

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

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

**SUMMARY OF THE DATA SET**

The movies data set comprises of 4803 rows with values entered for each of the 20 columns i.e. budget, genres, homepage, id, keywords, original language, original title, overview, popularity, production companies, production countries, release date, revenue, runtime, spoken languages, status, tagline, title, vote average and vote count.

**DATA CLEANING**

As part of the data cleansing process, code was written for the following columns to be removed as they were not considered when planning the analysis of the data set: 'keywords', 'homepage', 'status', 'tagline', 'original\_language', 'overview', 'production\_companies'. While this project coded for removing duplicate rows, after running the code, the dataset was still left with 4803 rows, hence the dataset had no duplicate rows. Movies with 0 revenue and/or 0 budget recorded were discarded. The release date column was converted into Date format to allow the the year from the date to be extracted. The budget and revenue columns were changed to integer format using numpy’s int64 method to allow performing mathematical operations. The values from the columns: genres, keywords, production\_companies, production\_countries, spoken\_languages were converted from JSON format into lists for each unique key name.

**MISSING DATA**

The information about the data frame function showed there was no missing data for the columns to be used in the analysis. Hence, no further rows- that would have had at least one missing value- had to be dropped.

**DATA STORIES AND VISUALISATIONS**

The five most expensive movies are presented below in descending order:

Graphical user interface, application

Description automatically generated

The five least expensive movies are shown in ascending order, with the cheapest in the first row, as follows:

Graphical user interface, application

Description automatically generated

The Modern Times movies stands out as the most profitable out of the five cheapest movies from the data set, with significantly high revenue and lowest budget. However, Modern Times does not make the list of the top 5 most profitable movies, shown in the figure below:

Graphical user interface, text, application, email

Description automatically generated

On the other hand, Avatar was the least profitable out of all 4803 movie entries.

Graphical user interface, text, application, email

Description automatically generated

The most popular movies were:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated 637 movies with votes above 7.

Further analysis showed that 1997 was the most profitable year for movies and drama was the most frequent movie genre, while foreign was the least.

A picture containing bird, aquatic bird, screenshot

Description automatically generated

Total profits increased overall from 1920 (first recorded date) to 2020, reaching two peaks shortly after one another around halfway through 2000-2020 and starting to decrease steeply after the second peak.   
Shape

Description automatically generated with medium confidence

However, for foreign movies, total profits were about constant from 1920-2000 with two exceptions: one trough after 1920 and one peak after 1980. However, total profits increased through 2000-2020 but they were volatile, nevertheless overcoming the peak after 1980 and reaching a trough around 2010.

Chart, histogram

Description automatically generated

To justify the increase in profits, the visualisation below shows that as more content is generated as time goes on, profits increase with the number of contents.

A picture containing shape

Description automatically generated

**THIS REPORT WAS WRITTEN BY: ROBERTA DOBRA**

**STUDENT NUMBER: RA23010006214**

