

**TASK**

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

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

**Introduction**

Dataset of movies including information on budget, release date, production country/ies, revenue, popularity rating, and revenue.

**DATA CLEANING**

First redundant columns were removed. These were: *homepage, keywords, id, status, tagline, original language, overview, production companies,* and *original title*. These were not deemed useful for the purposes of this EDA and mostly contained linguistic data, rather than numerical, which would not be useful for graphical visualisations. Further, *original language* and *original title* are both duplicated/expanded on in retained columns *spoken languages* and *title*, respectively.

Once these were removed duplicates were also deleted, using *title* as a reference point as this is one piece of data that should be totally unique.

Next, rows with 0 value *budget* and *revenue* were also removed as this would skew data and therefore visualisations, as it indicates missing data.

New column *release year* was added in order to facilitate further extrapolation and visualisation.

Columns *genres, spoken languages* and *production countries* were cleaned to be properly readable.

During each stage of the data cleaning process data samples were output in table format to examine usefulness of the newly cleaned data and inform further cleaning steps.

**MISSING DATA**

N/A values were not extracted, but as mentioned above, 0 value *budget* and *revenue* columns were removed.

**DATA STORIES AND VISUALISATIONS**

Table

Description automatically generated with low confidence

The top 5 most expensive movies in this dataset are:

Pirates of the Caribbean: On Stranger Tides, Pirates of the Caribbean: At World's End, Avengers: Age of Ultron, Superman Returns, and Tangled.

All of the top 5 most expensive movies made a significant profit, as can be seen when looking at the figures in the *revenue* column. This suggests the budget was justified.

Table

Description automatically generated

In order to examine the top 5 most profitable movies, a new column (*profit*) was created by subtracting *budget* from *revenue.* This produced the above.

The top 5 most profitable movies are:

Avatar, Titanic, Jurassic World, Furious 7, and The Avengers.

None of the top 5 most expensive movies are included on the top 5 most profitable movies, suggesting bigger spending does not necessarily equate to a larger profit.

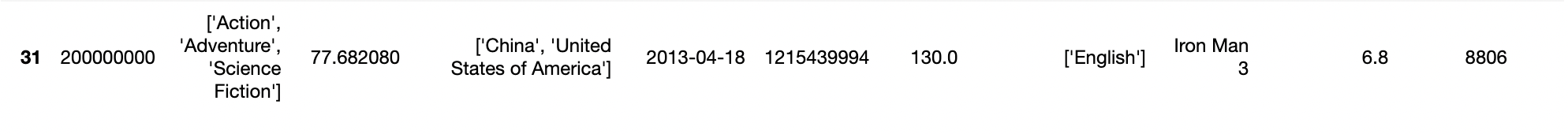
The *vote average* and *vote count* for the top 5 most profitable movies are also all higher than the top 5 most expensive.

The most profitable movie was released in 2009. Three out of the top 5 most profitable movies were released after 2010. The remaining top 5 most profitable movie was released before 2000.

Therefore, there is a gap from 1998 - 2008 where movies were not as profitable as these 5 movies.

Graphical user interface, table

Description automatically generated



By expanding the data to top 10 movies, it's clear that “Avatar” is an outlier as all other top 10 profitable movies were released from 2012 onwards. It is unsurprising that there is otherwise a gap until 2012 considering the global financial crisis in 2008, which will have both impacted studios' budgets and customers' spending power.

Table

Description automatically generated

The top 5 most popular movies, as extracted by *popularity,* feature on neither the top 5 most expensive, nor the top 5 most profitable. This suggests bigger budgets do not always guarantee views. It could also suggest that *popularity* is not measured by a finance-bound metric (i.e. ticket sales). If more ticket sales equals more profit, but higher *popularity* does not equate to higher profit, then *popularity* is not determined by ticket sales.

**Table

Description automatically generated**

In the final cleansed version of this dataset there are 3229 rows, of these only 637 are rated 7 or above. Of the 2 top rated movies, one (“There Goes My Baby”) has a popularity score that belies its vote average. This could render either/both *popularity* and *vote average* redundant, and further investigation into their method of measurement should be considered.

For example, if popularity is measured by number of viewers within a given period, does the result for “There Goes My Baby” suggest that despite not many people seeing the movie, the ones who did and voted, rated it very highly? Could it simply be a case of missing data? Especially when considering there were supposedly only 2 voters for “There Goes My Baby” in *vote count*.

Further, none of the top 5 most popular movies feature in the top 5 of the above output – which is sorted according to *vote average* in descending order. This could imply that even though a movie is popular, it does not guarantee voting action from the viewers, which also potentially renders *vote average* useless data, as it may only reflect a small subset of total viewers, and therefore a minority evaluation.

**Chart, bar chart

Description automatically generated**

The above bar graph shows number of movies per genre. “Drama” is the most common genre by quite a significant margin, whereas “Foreign” barely registers on the very lowest end of the scale. This is quite easily explained by the fact that a movie is only ‘foreign’ to its audience, and does not really constitute as a genre, given that it is entirely subjective depending on the language of the film, and who is watching it and where. A Spanish-language film is neither foreign in Spain, nor to a Spanish speaker in Wales, but it would be to a non-Spanish speaker in Taiwan.

What is interesting to note from this graph is that “Family” and “Animation” are not highly common genres, yet back in the top 5 most popular, the **most** popular film was “Minions” – ostensibly an animated family film. This could be because there is overlap in the genre classification and many films are considered multi-genre. For instance, “Minions” is labelled “Family”, “Animation”, “Adventure” **and** “Comedy”, which is the second most common genre. Further, of the other films on the top 5 most popular list only one (“Interstellar”) is classed as “Drama”, which is the most common genre according to the above. This implies that despite it being the most common genre, it is not necessarily the most popular.

**Graphical user interface, application, table, Excel

Description automatically generated**

The above histogram shows number of movies per production country. Immediately obvious is the dominance of the United States of America at nearly all 3229 movies that remained in the cleansed dataset. This is exponentially greater than all other production countries listed. This could be due to several reasons. Firstly, depending on the criteria for data selection, this could result in an over-representation of American-made films, and under-representation of other production countries. This is supported by the fact that it is also evident not all countries are included on the list. For example, there is no data for Nigeria, which famously has a huge output of locally made movies dubbed “Nollywood” which reportedly rivals Bollywood – itself an Indian output rival to Hollywood in the USA. Looking at the results for India on the above also suggests missing data, possibly in the cleaning process or the selection process.

A picture containing graphical user interface

Description automatically generated

**Background pattern

Description automatically generated with medium confidence**

**A picture containing graphical user interface

Description automatically generated**

The above bar chart illustrates number of spoken languages in all films. Unsurprisingly, given the dominance of American-made films already seen, English is the most recorded spoken language in this dataset. Beyond that, European languages also outnumber other languages.

**Chart, pie chart

Description automatically generated**

The above pie chart demonstrates proportion of film production per year. Most of the films in this dataset were made from 1998 – 2015. Likewise, there is little difference between the years, with 2006, 2009, 2010, 2011 and 2013 having the same level of movie production, and 2001 – 2005, 2007, 2008, 2012, 2014, 2015, and 2022 also having the same level of film production. Considering the above visualisations on production country and languages, we can also assume that each year most of the films made were produced in English, in America. The even spread of film production across years also suggests that the United States has a robust film industry that continuously makes new movies, as opposed to having peaks and troughs in their output.

Overall, from this EDA, it is evident that this dataset needs to be further cleaned in order to glean useful insights. Certainly, this dataset would not provide accurate analysis of global film production levels – as it does not include all nations. It also does nothing to illustrate audience sizes for different languages, as the dataset is heavily skewed towards English. Given these data holes, it is also reasonably safe to assume that were an analysis based on yearly production to be undertaken, this dataset would not include all films for a given period.

It could, however, potentially provide useful insights for analysis of American-made, English language movies specifically. Therefore the data would need to be further cleaned to only include those parameters. Also, further investigation into how things like *popularity* are measured would also be needed to make accurate conclusions thereafter.

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

