

**TASK**

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

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

**Introduction**

The movie data set is comprised of categorical and numerical features on 4806 movies produced and released over the 100-year period from 1916 to 2017.

The most notable numerical features for each movie are, the budget, popularity, revenue, vote count and average. The most notable categorical features are genres, production countries, spoken languages and title.

**DATA CLEANING**

The first step of the data cleaning process comprised of identifying columns that are redundant or unnecessary namely: homage, keywords, original\_language, original\_title, overview, production\_companies, status and tagline.

The second crucial step in the data cleaning involved the identification and removal of the duplicated movie entries (rows) in the dataset.

The last step of the data cleaning involved converting the *genres*, *production\_countries*, and *spoken\_languages* categorical features, from the JSON format to a list of strings.

**MISSING DATA**

The data set had missing values, empty cells, and entries with zero budget/revenue.

Due to the relatively large size of the data set, the above-mentioned entries were removed from the dataset.

**DATA STORIES AND VISUALISATIONS**

The most significant insights are gained from the visualisations of the categorical movie genre feature and some key numerical features.

The most notable observation is that budget does not necessarily translate to profitability, as the top 5 most expensive movies are not in the top 5 most profitable movies.

Similarly, highly rated movies by establishment platforms, are not necessarily the most popular among the movie goers.

The movie genres of Drama, Comedy, Action, Agriculture and Horror have the highest frequency of movies in each genre and are also the most popular movie genres in the dataset.

The main financial features (Budget, Revenue & Profit Earned) in the dataset have been steadily increasing over the release years.

There is notable correlation among these features (Budget & Revenue) with profits earned and are almost perfectly correlated to each other. The 3x3 Pair-Grid scatterplot shows that a movie classification model can be built based on these financial features.

The distribution in the diagonal’s points to a skewed gaussian distribution, which lands to standardised feature scaling for revenue and budget.

Furthermore, more insights can be gleaned from analysing popularity versus vote count/average. Popularity and **vote\_count** have a noticeable correlation based on the scatterplot. While popularity and **vote\_average** have a poor correlation as per the scatterplot, reinforcing the earlier mentioned observation.

However, for more advanced analysis the numerical features ought to be either normalise or standardised before modelling and advanced analytics are implemented.

# ENSURE THIS DOCUMENT IS NEAT AND ADD IT TO YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY: ZUKISA XOTYENI**

