Exploratory Data Analysis for Machine Learning - Report 1

Task Overview

For this report, I was tasked with picking out a dataset and performing various aspects of Exploratory Data Analysis and hypothesis testing. These included data cleaning, feature engineering and creating visualisations inside of a python notebook. The dataset that was used was the Latest Movie Dataset 2025 that can be found on Kaggle.

Initial Data Exploration

To begin, the data was loaded into a pandas DataFrame and various pandas methods were used to find summary statistics of the data. The data consists of 7 columns, which indicate the title, release date, original language, various voting scores and an overview of the film. The initial findings are as follows:

1. The dataset has 10000 rows and 7 columns.
2. All rows except popularity, vote_count and vote_average are strings with the popularity and vote_average being floats and vote_count being an int.
3. The release date and overview columns are the only ones containing missing values, with there being 49 and 242 missing values respectively.
4. The popularity score or vote_average score can be used as target variables for machine learning models to predict a movies success.

One surprising finding here was that the popularity scores might have some errors since the column has a mean value of 7.413149, but a maximum value of 1096.665400, with later entries into the dataset having much lower scores than the first ones, indicating there could have been some issues with its calculations.

Data Cleaning and Feature Engineering

To begin data cleaning, the release date column was converted into a pandas datetime object which in turn allows for easier manipulation and operations to be done. The next step was to add a release year and release month column which was derived from the release date which will help with creating visualisations in the future. Then the rows with missing values were dropped as they made up a very small proportion of the data and using an imputer is not appropriate for a release date or overview column.
Next the original language feature was encoded using a one hot encoder which will allow for models to extract certain patterns from the language of the movie. To further clean the data, all columns where the vote count was 0 were dropped as if they had no votes, they would not have a meaningful contribution to an average vote score.
Further feature engineering steps included converting population and vote count into their log versions to create easier to understand visualisations as can be seen in Figure 1

After applying log transformations, the popularity more closely resembles a normal distribution while the vote count is closer to a uniform distribution. This is not expected behaviour as it should still be closer to normal seeing as more popular movies should receive more votes with less popular

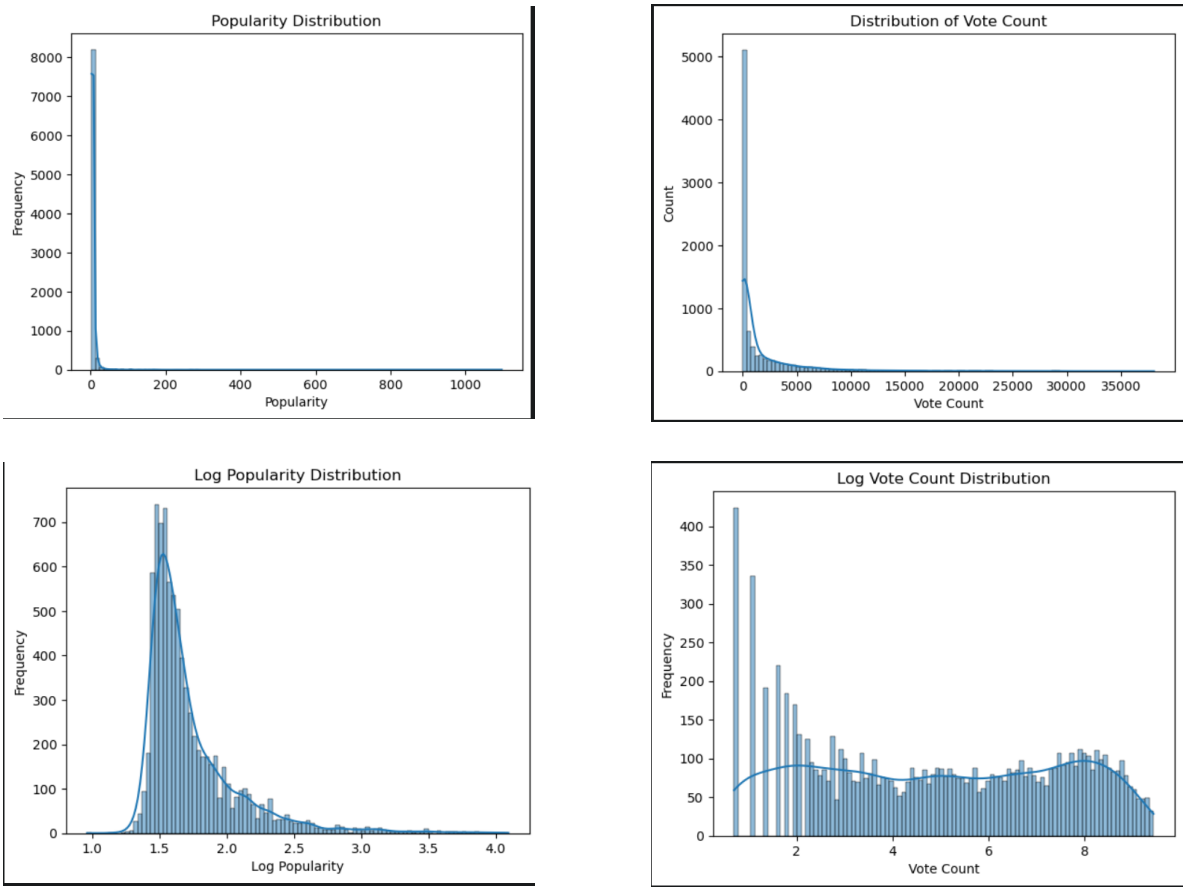

Figure 1: Difference between a distribution and its log counterpart

ones receiving less votes, meaning there may be some issues with the calculation of this or the dataset might have some bias.

On top of the log transformations, extreme outliers were also removed by the process of calculating their z score and dropping values whose absolute z score exceeded 3. This improved the visualisations further and dealt with some of the popularity outliers that were discussed earlier.

Key Findings and Insights

For this section we looked at the popularity overtime for movies, the popularity based on release month and the number of movies per language. In Figure 2 we see that the popularity of movies steadily increases over time with a large increase in 2020.

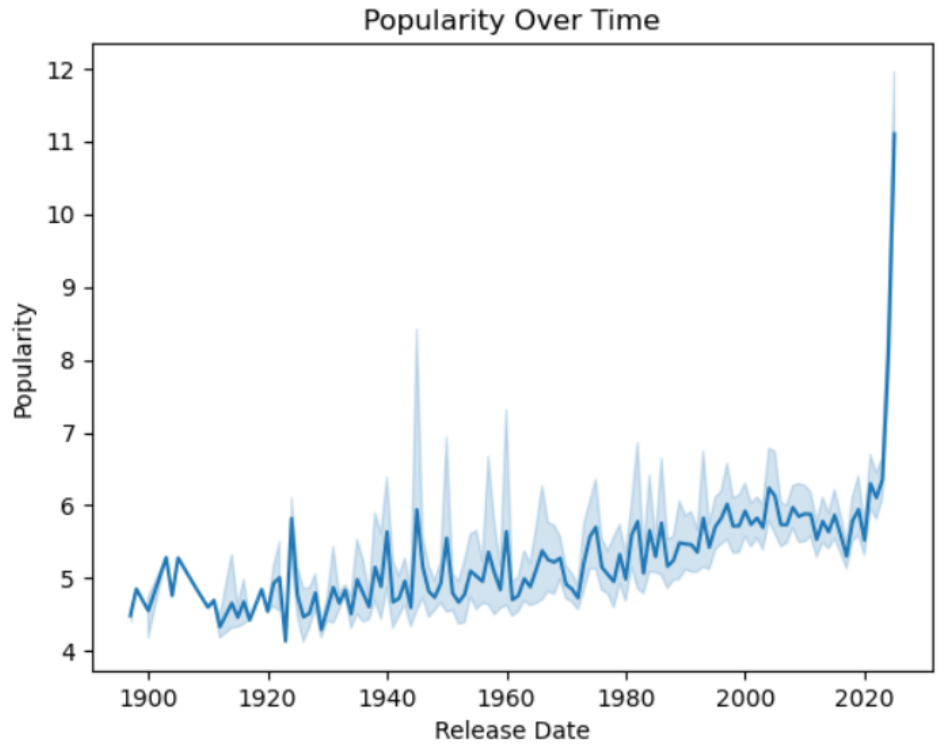

Figure 2: Popularity of Movies over time

This is to be expected as movies have become much more popular and accessible over time with a large increase around 2020 with the effects of COVID and more people watching a large range of movies.
In Figure 3, we see that the release month for movies does not vary much but the summer period does have more movies released that the other months by a marginal amount.

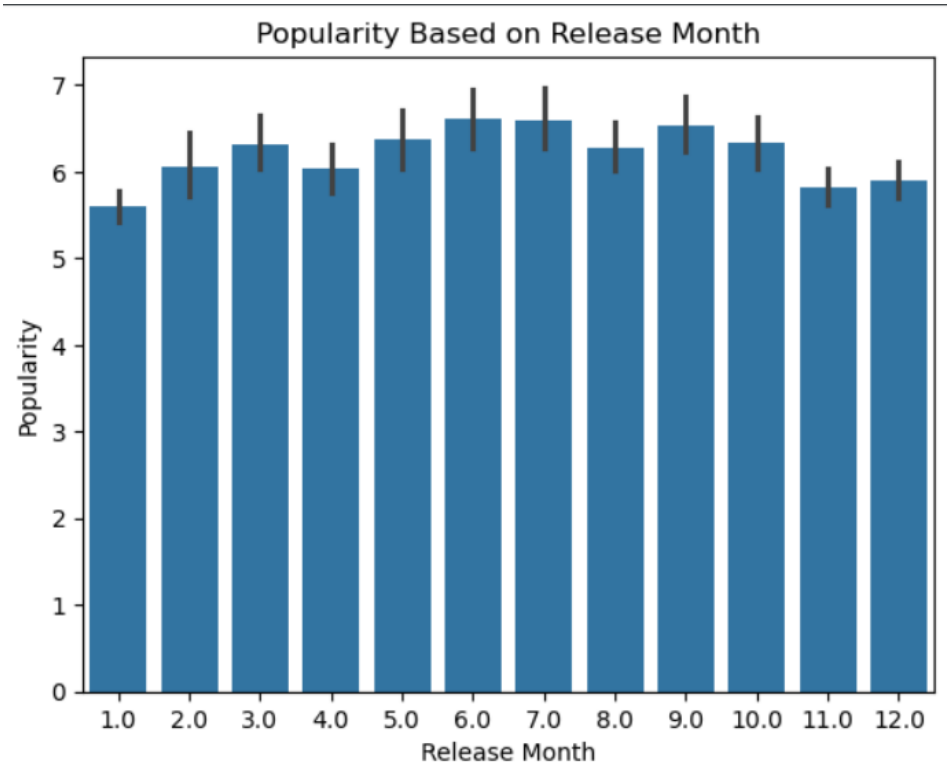

Figure 3: Amount of Movie Releases per Month

This is also to be expected, as during the summertime, more people are available for longer periods of time, especially in the case of university students or school kids, meaning releasing in that window could lead to more sales due to increased availability of customers.
Figure 4 shows the top 10 languages that have the most movies made about them. The graph shows English movies as the most numerous, with languages such as Japanese and French being much lower.

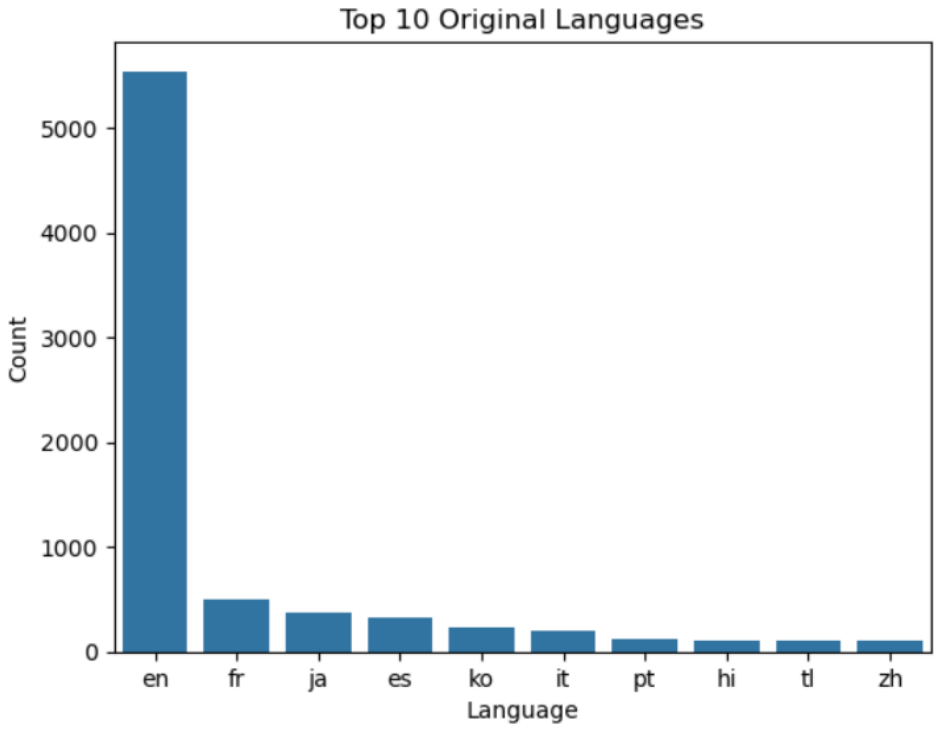

Figure 4: Top 10 Languages for Movies

This can be expected to a degree since most of the movies we consume are in English due to its vast global presence and the American Movie scene being large, however it is surprising to see such a low amount of Chinese and Korean movies with their recent rise in popularity and with how many movies are strictly made for China and its population. This is most likely an issue with the source of the data alongside the difficulty of obtaining data about Chinese movies and their popularity.

Hypothesis Testing

There were 3 main hypotheses that were tested in this section. All hypotheses were tested with an α value of 0.05.

Hypothesis 1

Null Hypothesis (H_0): Movies released after the year 2000 are *not* more popular than movies released before 2000.
Alternative Hypothesis (H_1): Movies released after the year 2000 are more popular than movies released before 2000.
A single tailed t-test was performed here which gave the results of:

$$t = 18.48, \quad p = 4.86 \times 10^{-75}$$

Since $p < \alpha$, we reject the null hypothesis and conclude that movies released after the year 2000 are significantly more popular than movies released before the year 2000.

Hypothesis 2

Null Hypothesis (H_0): English movies do *not* receive a higher vote average than non-English movies.
Alternative Hypothesis (H_1): English movies receive a higher vote average than non-English movies.
A single tailed t-test was again performed which gave the results of:

$$t = 1.49, \quad p = 0.0679$$

As $p > \alpha$ we fail to reject the null hypothesis and conclude that English movies do not receive a significantly higher vote average than non-English movies.

Hypothesis 3

Null Hypothesis (H_0): There is no difference in average popularity between movies released in different months.
Alternative Hypothesis (H_1): At least one month has a different average popularity.
A one-way ANOVA test was performed which gave the results of:

$$F = 7.896, \quad p = 7.90 \times 10^{-7}$$

As $p < \alpha$ we reject the null hypothesis and conclude that there is a significant difference in average movie popularity across different months.
This result is surprising a there does not seem to be much variation in the visualisation however, the release month does play a part in the popularity of the movie. To further research this, we can look into which month has the largest popularity and what factors lead to this. This information can further help guide decisions into when to release a certain movie.

Key Takeaways and Future Steps

The initial data has been thoroughly cleaned and prepared for more advanced tasks to take place. This data can be used to create a predictive model for a movies popularity based on factors such as language and release date. Further analysis can be done on the overview of the movie to extract certain keywords that explain the genre of the movie or the mood/setting to create further features. Although this dataset is quite simplistic, a lot can be done to make it useful and applicable to future movies.
Key takeaways include:

1. The release month of the movie is statistically significant.
2. Movies have become much more popular over time.
3. More data may need to be added to create reliable predictive models for the success of a movie.