The Movie Database (TMDb Analysis)

The data set I worked on was the movie database data set which consisted of 10,000 movies and data collected on them to be analysed.

The research question posed to be answered were:

- Which genres have been the most popular?
- Which properties are associated with higher revenue?

I loaded the dataset with the path to file and used a set option function to allow all columns to be displayed so I could wrangle the data accordingly. I explored the data set, looking into the number of columns, rows and used the info and describe function to familiarize myself with the data set, and also to figure out what I would eventually filter. I dropped all null rows. I changed one of the data types of two columns, budget_adj and revenue_adj that were in the form of exponential functions, to make it easier to work with.

To clean the data, I used the assign and explode function to transform the genres that were separated by pipes (|) in the genre column into individual rows, making them much easier to work with.

To better explore the first question posed, the mean of popularity grouped by the genre was found and analysed in a bar plot and a scatter plot. Also, the mean of the vote count of various movies was analysed along with the popularity.

To better explore the second question, the mean adjusted revenue and genre were compared in a scatter and bar plot. Patterns of the adjusted revenue and release year were also evaluated to find which properties are associated with higher revenue.

Findings

- Which genres have been the most popular?

In conclusion, I found three main genres that seem to be the most popular as well as have the highest vote counts. These three genres were Adventure, Fantasy and Science Fiction. The mean as well as the bar and scatter plots showed these patterns.

- Which properties are associated with higher revenue?

In conclusion, the most popular movie genres seem to be related to the movies that make the most revenue, shown in the bar and scatter plots. The years the movies were released show more revenue in a scatter plot. However, older movies have no information on their revenue and so this might be the reason for this reading from the scatter plot.

- Limitation(s)

Some records of adjusted revenue of older movies from the 1960s seem to be absent. This might lean the data towards the positive correlation between adjusted revenue and the release year in the second research question.

References

https://seaborn.pydata.org/tutorial/relational.html

https://pythonguides.com/matplotlib-rotate-tick-labels/

https://stackoverflow.com/questions/66779466/how-to-use-explode-function

https://www.geeksforgeeks.org/show-all-columns-of-pandas-dataframe-in-jupyter-notebook/

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.bar.html