# Project Report – James Ahern

## GitHub URL

https://github.com/jahern2/UCDPA_JamesAhern

## Abstract

This project is intended to serve as an analysis of a dataset which contains large amounts of information about the HBO Max streaming platform. HBO Max is a relatively new streaming platform and was launched in May 2020, with the dataset including data up to May 2022. With this being a relatively new streaming platform I was curious to see how much of the content was released recently, how much older content they have on the service and what other interesting insights I could discover by applying the various skills that I have learned throughout this course.

I am also excited about the prospect of the platform launching soon where I live and in other European countries as I watch a lot of HBO content.

## Introduction

I had initially intended on using a dataset containing Playstation 4 games and merging that with a Metacritic ratings dataset and analysing that but the Playstation 4 datasets that I found had a lot of issues with missing values and as a result, I wasn't able to get many valuable insights from the data.

I then pivoted to this dataset for a few reasons - I watch a lot of HBO content so it is interesting to me, I am excited about the prospect of the platform launching soon where I live and I think that I can get some valuable insights from the dataset.

## Dataset

The dataset that I chose was a HBO Max dataset that I found on Kaggle (HBO Max TV Shows and Movies, 2022). The dataset contains two csv files which was useful for demonstrating what I have learned about merging dataframes. I had initially intended on merging the dataset with an IMDB dataset but IMDB ratings and other info were already contained in the main csv file. I chose this specific dataset because the person who uploaded it has many other highly rated datasets with a large number of downloads on the site.

The dataset contains information about all content available on the HBO Max streaming platform as of May 2022

The main dataset has 3294 rows and 15 columns, this is more columns than I would like and I didn't need a number of them so I removed some before merging with the second csv file that accompanied the main file, more details of this will be included in the Implementation Process section.

# Implementation Process

**Importing Data:**

- The first steps of my process were to import the packages that I would use throughout the process, these being numpy, pandas, matplotlib and seaborn, and the main dataset that I would be using for the project. I set the index column to 0 so that it reflected the "id" column on the dataset, I did this because I knew that I would be merging with the second dataset on this column. The first dataset contained information such as the movie/tv show name, imdb rating, release year etc.

**Preparing the datasets and merging:**

- I first previewed my dataset using .head and noticed that there were a few things I needed to clean up such as the column titles, I then checked for duplicate entries of which there were none and missing values. There were large amounts of missing values in two columns but thankfully they weren't columns that I was planning on using.
- I then made a copy of my dataset, proceeded to drop several columns and removed unnecessary square brackets and quotation marks from the rows throughout two columns. Lastly, I renamed all columns to make the dataframe look better.
- With my first dataframe looking good I then imported my second dataset, which contained actor/director names and character names for the movies and tv shows in the first dataset. I made a copy and cleaned it up in a similar way to the first dataset but in this case, I subsetted the roles column (which included actors and directors) using .loc so that I was left with just a list of directors. This reduced the number of rows from 66,393 to 2,774 which is much closer to the number of rows in the first dataset.
- I was then satisfied that both dataframes were cleaned and prepared. I proceeded to merge the dataframes on the id column using the .merge function. This was successful and I was happy with what I saw so I proceeded to the next step.

**Sub-setting the dataset to create some interesting charts:**

- My next task was to create a dataframe of the top 10 directors with the most director credits on the HBO Max platform, I achieved this by using the .value_counts function and adding the values to two columns – 'Director' and 'Count', I ended up including 11 directors since there were 4 directors with 10 movies each.
- I then created another dataframe containing the top 3 countries based on where content on the platform was produced, these two dataframes were utilised in the next section when I visualised the data.

**Visualising the data using charts and graphs:**

- The main tool that I used for visualisation was Seaborn (Example gallery — seaborn, 2022). I used this because I was more comfortable using this tool than matplotlib and I thought it handled the dataframes that I created better.
- I created various charts such as histograms, horizontal bar charts and line plots. I changed up the colour and style throughout with the intention of making my charts more engaging to the viewer.

## Results

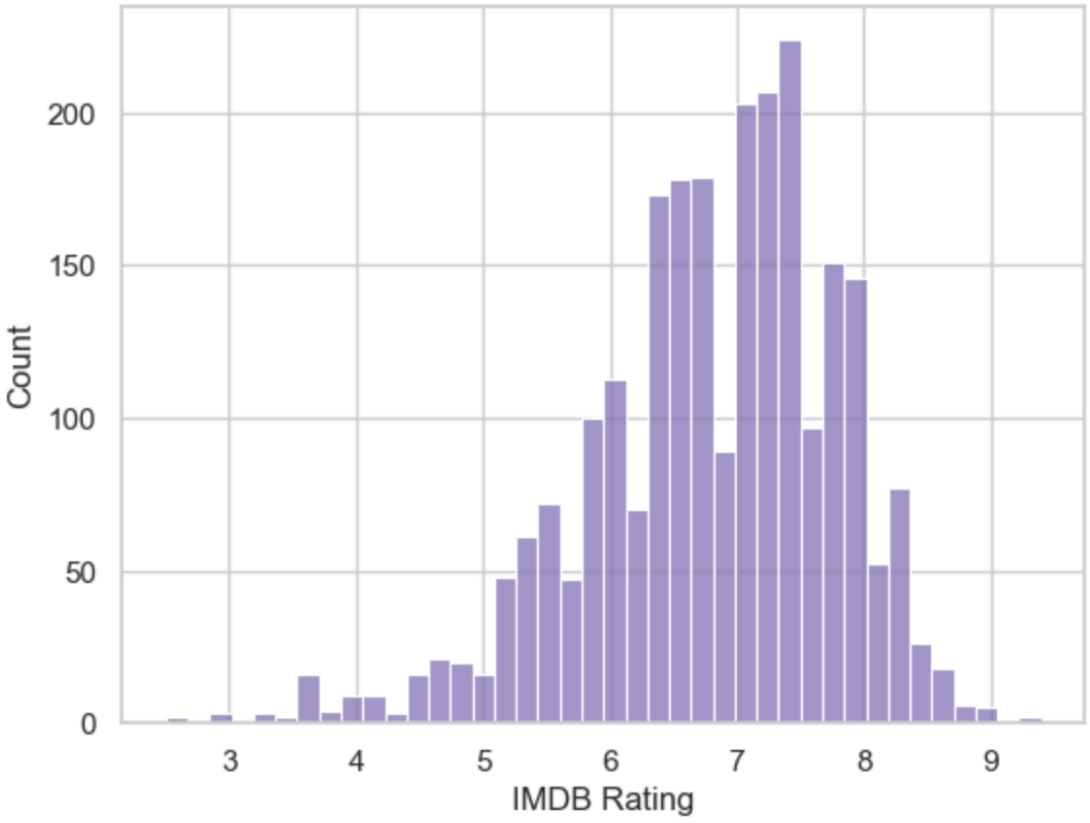


*Figure 1 - Histogram showing the IMDB rating of movies and tv shows on the platform*

In this chart we can see the count of IMDB ratings for content on the platform, most of the content falls between 6.0 and 8.0 with 7.0 to 7.5 being particularly common ratings.

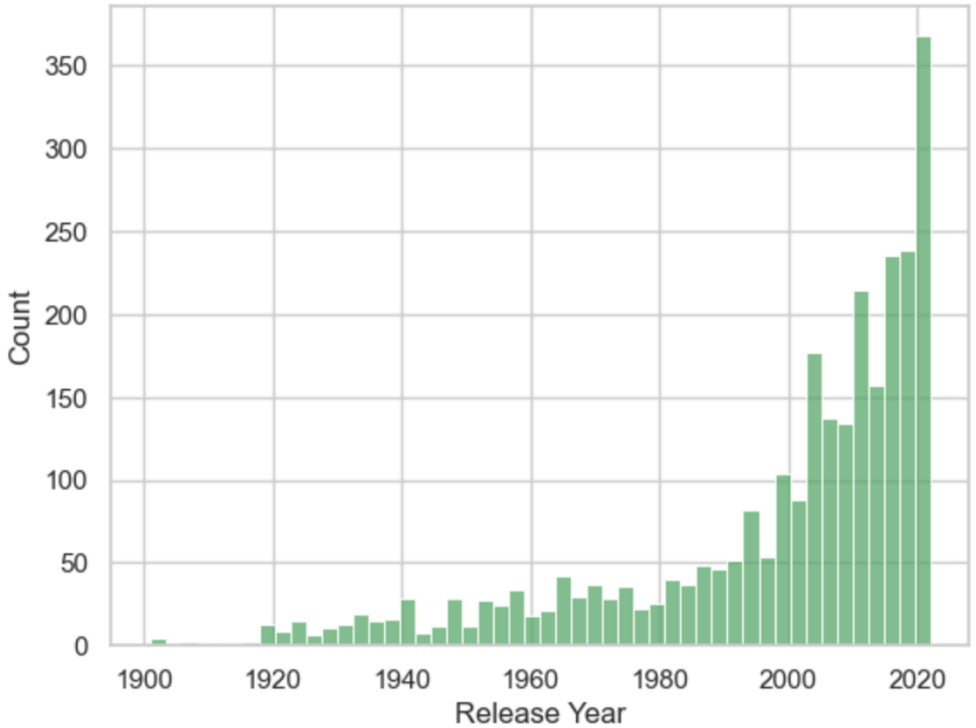


*Figure 2 - Histogram showing the release year of content on the platform over time*

This histogram shows the release year of the content on HBO Max. With this being a new

platform, it is not a surprise that there is a lot of new content from 2021 and 2022. It would be interesting to see how this compares to an older streaming platform such as Netflix. I was quite surprised at the amount of older content that was available on the platform but with this being HBO they have quite a large back catalogue of movies and tv shows to draw from.

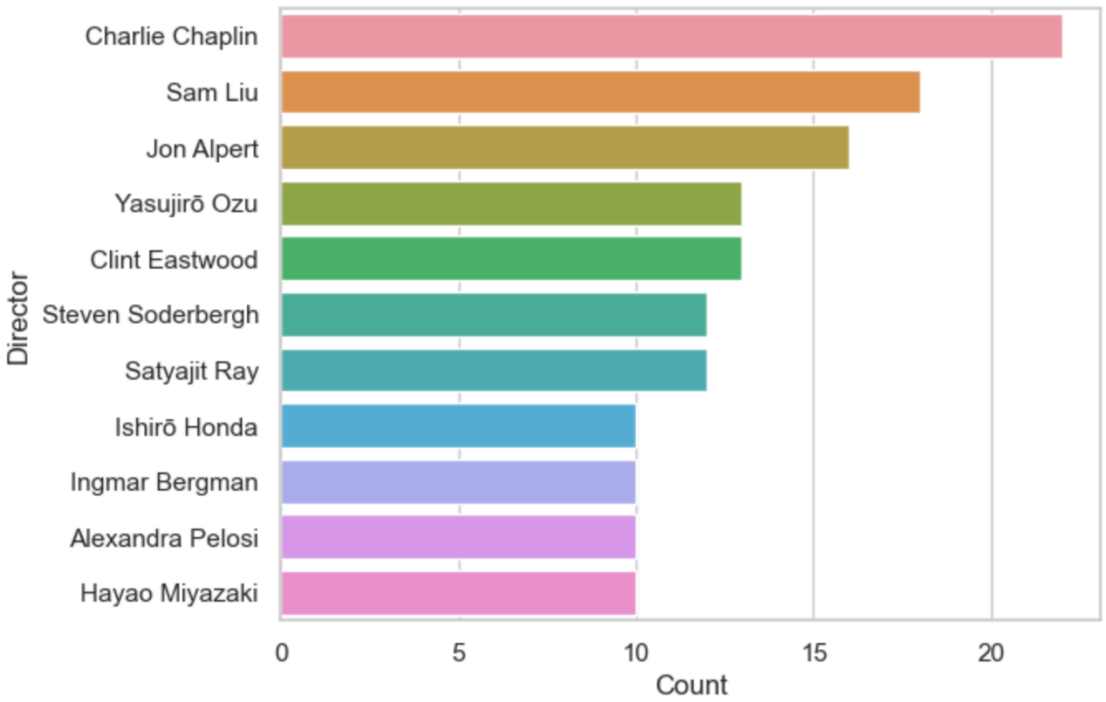


*Figure 3 - Horizontal bar chart showing the 11 directors with the most director credits on HBO Max*

This horizontal bar chart which shows the directors with the most credits on the platform was my personal favourite, Charlie Chaplin is listed as the director for 22 movies on the platform with some other household names such as Clint Eastwood also featuring on the list with 13 films to his name. It is interesting to note that Chaplin also featured as an actor in all the films that he directed which would be very uncommon these days, with him playing roles such as Henri Verdoux in his film Monsieur Verdoux which is about an unemployed banker.

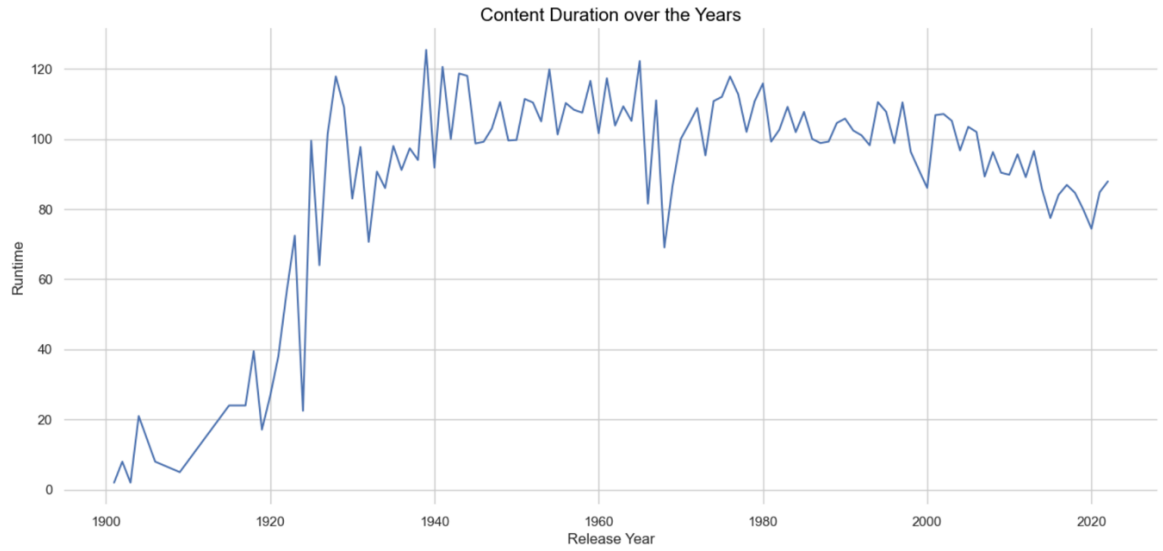


*Figure 4 - Line plot showing the average runtime of content on the platform by release year*

This line plot illustrates the average runtime of content on the platform over time, you can see that the runtime remained relatively stable between 1940 and 2000 at 100-120 minutes but it has since fallen to 80-90 minutes. It is relevant to note that I had originally created a scatterplot (which I didn't end up using) that showed a large increase in content in the 30-45 minute range over the last 5-10 years, this content is likely mostly tv shows and may be part of the reason that the content duration has fallen in recent years.

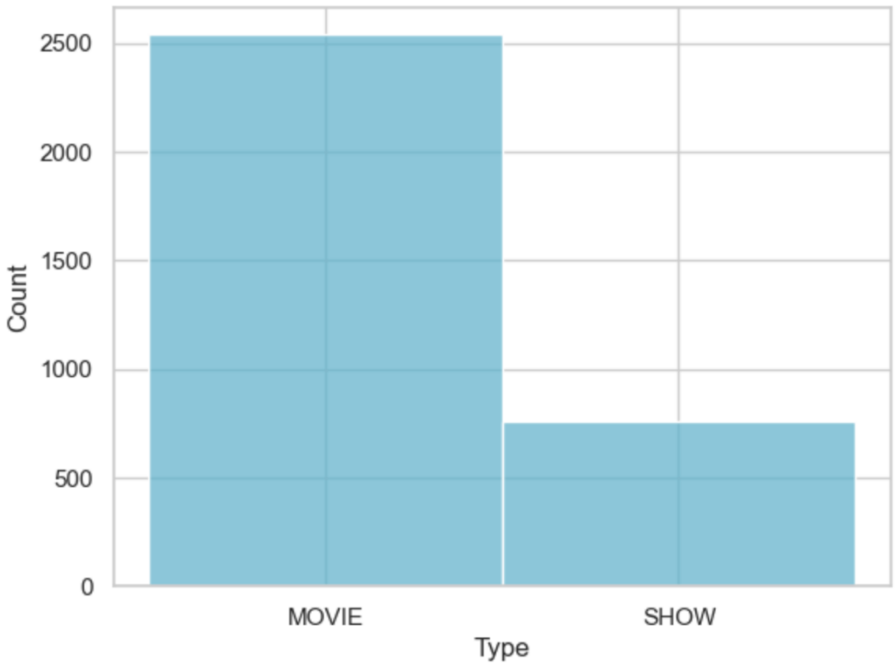


*Figure 5 - Bar chart showing the amount of Movies compared to the amount of TV Shows on the platform*

The bar chart above shows the number of movies in comparison to the number of TV shows on the platform, as you can see there are much more movies than TV shows at present.

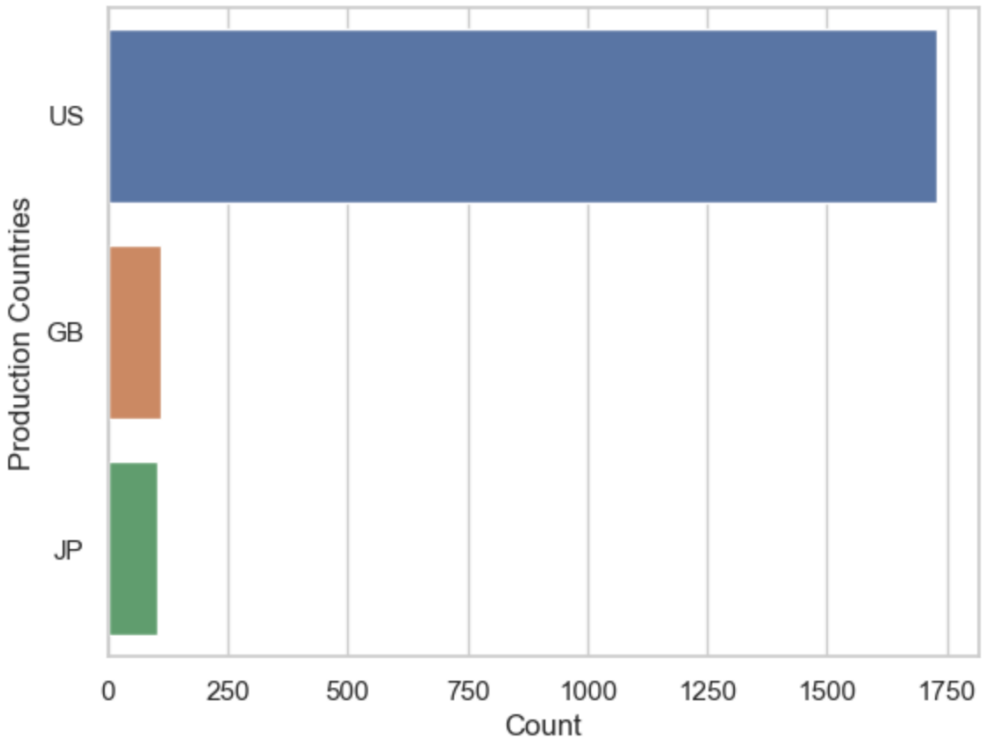


*Figure 6 - Chart showing the top 3 countries where content on the platform was produced*

This chart shows the top 3 countries where content on the platform was produced, as you can see the United States dominates in this category, with Great Britain and Japan in second and third, however, it is important to note that this data was collected from content available in the United States on the HBO Max platform. It would be interesting to see how this data would change (or wouldn't change) if it was collected from another country where the streaming service is available, such as Portugal or Brazil.

## Insights

- Much of the content on HBO Max has been released in recent years, with over 350 of the movies and tv shows on the platform being released in 2021 and 2022, this makes up over 10% of the content.
- Despite HBO having a reputation for creating popular and highly rated movies and tv shows the IMDB ratings suggest that there is a lot of poorly reviewed content on the platform too, with more content being rated between 5.0-6.0 than 8.0-9.0 on IMDB.
- Charlie Chaplin has the most director credits on the platform, even though the first movie he directed was released in 1914.
- The runtime of content on the platform based on release year has fallen consistently since around the year 2000.
- Most of the content on HBO Max was produced in the United States, with 1,730 movies and tv shows being produced exclusively in this country, the next two countries on the list have significantly less with Great Britain producing 109 movies and tv shows and Japan producing 102.

## Machine Learning

If I was to make predictions based on this dataset using machine learning and/or deep learning some interesting predictions to make could be predicting how the runtime of content will change over time and/or predicting how the ratio of movies to tv shows will change over time. This would be interesting because of the greater number of tv shows that were released in recent years on the platform.
If I was to use machine learning to make predictions based on this dataset I would use regression methods because I would be predicting a number (eg. 60 minutes is the expected runtime for content in 2030), and not categorising data into classes or categories.

## References

Kaggle.com. 2022. HBO Max TV Shows and Movies. [online] Available at: https://www.kaggle.com/datasets/victorsoeiro/hbo-max-tv-shows-and-movies [Accessed 22 September 2022].

Seaborn.pydata.org. 2022. Example gallery — seaborn. [online] Available at: https://seaborn.pydata.org/examples/index.html [Accessed 22 September 2022].