Project Report – Mark Keogh

GitHub URL:

<https://github.com/mkeogh-ire/-UCDPA_MarkKeogh>

Abstract:

The project represented in the above GitHub repository is an analysis of streaming platform Disney Plus. The dataset contains information, such as original release year, director, title etc, on the Movies and TV Shows which were on Disney Plus as of mid-2021.

Using Jupyter Notebook, python 3 environment with modules imported, we are able to breakdown the dataset to give use information relating to aspects of Disney’s streaming content.

The hope of the analysis is to analyse the content of Disney Plus to determine if there is a bias to certain countries or languages. Whether the majority content is suitable to the family orientated reputation of Disney or are they branching towards a more mature audience set. Considering the success Netflix has seen over the last few years and their reputation for hard hitting, edgy, mature content they produce and upload, is Disney competing against the same target audience or are the sticking to the family friendly audience which has made them such a success.

Introduction:

The reason for choosing this project use case was for a couple of reasons. Firstly, I thought it would best represent the tools and skills learnt throughout the course. Secondly, I have a Disney Plus account with a young child at home so it’s interesting to me to see what variety of shows they may be susceptible to watching whether intentionally or not. Also given the large variety of datasets available through Kaggle, Google etc I thought this was most interesting to me. Ideally, I would have preferred to use a dataset from work however GDPR prevents me from doing so.

I felt that because I have a Disney Plus account and actively use it, this dataset presented as a real-world scenario for me to use.

It is also interesting to see from the data if the streaming service is all inclusive, multiple languages and different countries providing the content from them.

Dataset:

The dataset itself, original csv file saved to GitHub repository, is as stated a comma-separated values file.

The first row is the heading of the data, each column named to refer to what data is stored in the dataset. Here we have show\_id, type, title, director, cast, country, date\_added, release\_year, rating, duration, listed\_in, description, stream\_platform.

The data contains 1450 rows of information. Show ID is a unique identifier for each movie or TV show. Type is either Movie or TV Show. Title, director, cast and country speak for themselves. Date\_added is the date it was added to streaming service, release\_year is contents initially release year. Rating is the Parental advisory rating associated with content. Listed\_in is the genre of the content, description is a line about the movie and stream platform is self-explanatory.

Implementation Process:

I broke down the process into different headings which will be described below.

Importing the data

First step of the process was to import our different packages to be used in the analysis. Here we imported numpy as np, pandas as pd, matplotlib.pyplot as plt and seaborn as sns.

Next, I import my csv file to Jupyter using pd.read\_csv naming the data frame as df. Using .copy() to take a copy of the original dataset.

Continuing on, I then assessed the dataset using .shape() to find the number of rows and columns. Then using .head() I could visualise the first 5 rows. Next using .info() to find column types and null values and lastly using .unique() which tells me number of unique values in each column.

Cleaning the data

Using .astype(‘datetime64’) I converted the ‘date\_added’ field to a proper date and time format this allows for better analysis of the data withing that column as previous data in field was in November, 12, 2001 format. Next I checked for null values using the isnull().sum() method. This returns the null values in each column. From this information I could see director, cast and country were the only columns with null values present. Determining the director and cast columns were not need I remove them using .drop(column={‘ ‘}). Then using .fillna I was able to replace the null values in country column with ‘Unknown’. I renamed the ‘rating’ column to make it more readable using .rename.

I then converted the ‘type’ column to a category using astype to better be able to read the data when the time came to chart it.

Check for duplicates in data.

I then created a new column for month added which pulled the month from date\_added field and convert it to show only the month. This will again allow for a better analysis of the fields. This was done using .strftime and in order to do this I had to import datetime from datetime.

I also created a column for Year added, pulling this information from date\_added again, using .apply.

I then dropped the date\_added field.

Created a list for the rating, to make the categories easier for end user to read. Once list was created, I then matched these values to the values in rating and created a new column n called ViewerAge. Which I then categorised.

For the country column, a lot of the values were of multiple countries meaning the data would be hard to analyse so to workaround this I manipulated the fields to only show the first country in the value. To do this I used .apply function and split the strings to then only show the first value. i.e., field one contained United States, England, India and after it only showed United States.

I then split the dataframe into two dataframes, Movies\_df and TVShows\_df. This was to help when comparing the two.

Analysis

First analysis I done was by type, Movie and TV Shows. I got the shape of the column which showed the correct, 1450, rows. Then I got the value counts, 1052 for movies and 398 for TV Shows. Represented this as a bar chart and then as a pie chart for percentages, 73% for movies and 27% for TV Shows.

Next, I done a similar analysis on the Viewer Age field. This gave a breakdown in percentage of the tv shows and movies directed at kids, teens or older kids.

Continuing I ran a histogram on the release to with bins of 30, a distribution plot and boxplot on the release year of each movie or tv show. The box plot will give us the minimum year, the first quartile, median, third quartile and maximum value. I used the whitegrid style on the boxplot. Seaborn has a number of pre-built styles to choose from.

From here I ran a bar chart based on the top 20 years for which the content was initially release.

Also looking at the count for year added which I did for Movies and TV Shows separately.

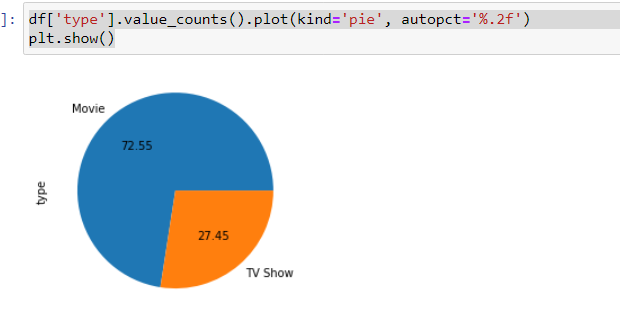
Next, I look at a scatter plot comparing release year to country with a hue based on the type of column.

Using seaborn again I then ran a pair plot on release year and yearadded. This separated Movies and TV Shows.

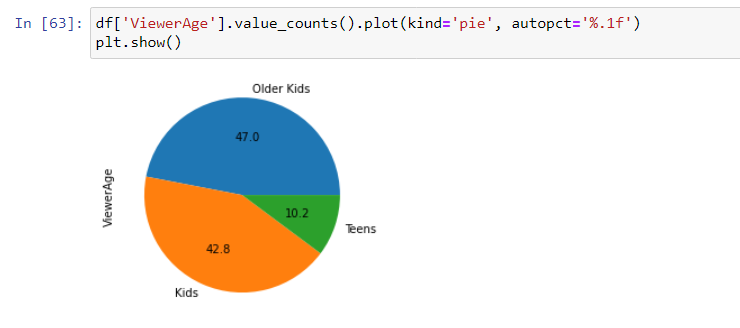
Lastly, I grouped release year. This gave a count of how many shows were release each year. To do this I had to reset the index. I ran this as line plot and then as a reg plot to find the regression line.

Results:

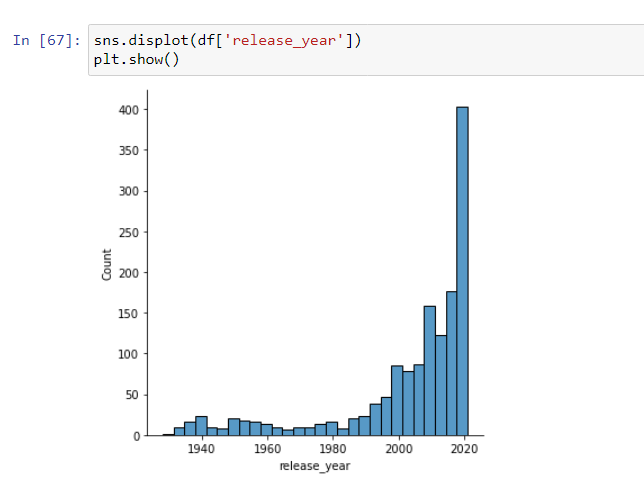
Results from these different analysis are as follows:



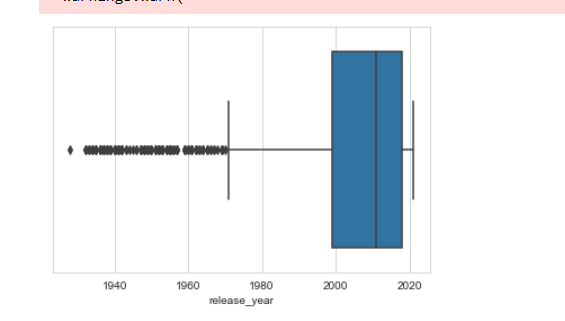
This shows that the volume of Movies is considerably higher than that of TV Shows, 73% to 27%.



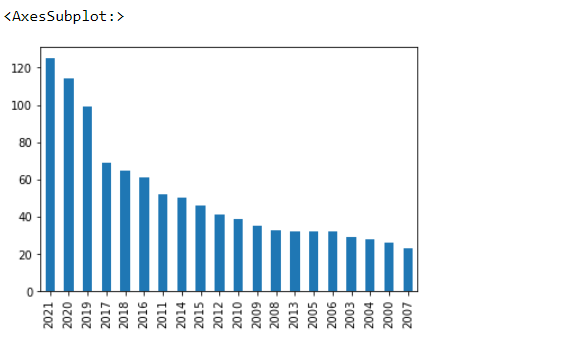
Here we can see that most of the content is intended for kids and older kids, with a small fraction intended for teens. 47& for older kids, 43% for kids and 10% for teens.



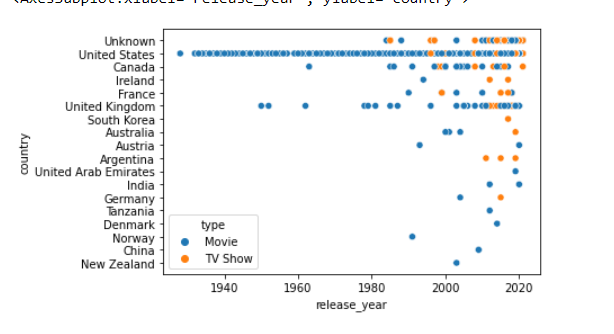
Next, we can see the majority of content had a release year of 2020 plus. Not highly insightful but shows that Disney is intended on release new content more than older content.



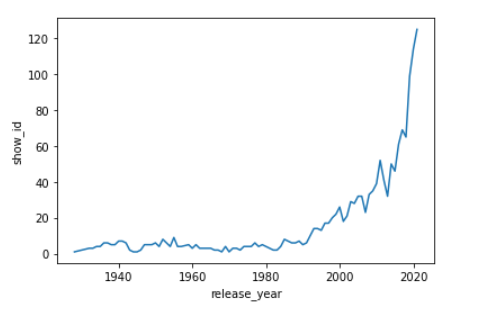
Here we see the box plot for release year, with the minimum year being 1928 (oldest film), mean being 2003, first quartile being 1999, third quartile being 2018 and the newest film being 2021.



Here we see the top 20 years for film releases, 2021 having the films followed by 2020.

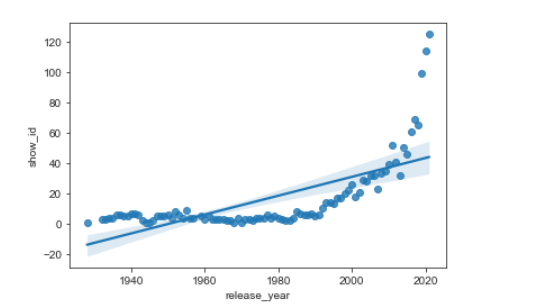


Here we can see the release year compared to country of origin. We can clearly see a bias towards United States films however it seems more films are being added from non-English speaking countries.



From here we can see the number of shows released based on the year. Can see a sharp incline from 2019 onwards, this is because Disney plus became live in 2019.

Lastly, we have the regression plot, this gives us an of the release year moving forward.



We can see that Disney is more inclined to continue releasing more newer content going forward.

Insights:

* Most content on the streaming service Disney plus is from English speaking countries.
* Content is favourable intended for kids and older kids
* Disney is sticking to its reputation of a family orientated company
* They will continue a trend of releasing more new content instead of adding more older content compared to new.
* Not surprisingly the top 3 years for content added are 2021,2020 and 2019. Considering Disney Plus went live in 2019 this makes sense.

References:

Dataset downloaded from here:

<https://www.kaggle.com/datasets/shivamb/netflix-shows>