Netflix TV Shows and Movies
Exploratory Analysis

Langdon Hatton, lhatton@bellarmine.edu
Leighann Robinson, lrobinson4@bellarmine.edu

I. INTRODUCTION

In the mainstream world of today, gone is the major usage of television providers, and owning DVDs. Many individuals in the United States today have some form of streaming platform account, with the streaming service Netflix, being one of the most popular. Our mutual interest and appreciation of the platform directed us to further exploring some data from the service. The dataset used in our exploratory analysis is titled "Netflix TV Shows and Movies," and contains information regarding the shows and movies that can be found on Netflix as of July 2022. July 2022 is the end of the data range due to this being when the data was originally collected and made available to the public. The dataset was downloaded from the website Kaggle, which is an online community for data scientists and individuals interested in machine learning. Our dataset was highly rated and upvoted from members of the Kaggle community, offering us a reliable dataset to further analyze and explore using different means. In our exploration, we conducted our analysis using Tableau, and Python libraries Matplotlib, Pandas, and Seaborn.

II. DATA SET DESCRIPTION

The dataset that was used contains 5850 samples, stored in 15 columns with various data types. A complete listing of our variable names and their data types is shown in table 1 below. Table 1 also shows the percentage of data missing from those columns in the original dataset prior to the data cleaning we conducted. The column that had the most null values was the seasons column, however, we believe this is due to the dataset including both movies and television shows, and of those two, movies do not have seasons. To counteract this, these null values were changed to "No Data" in order to satisfy the title being of a movie or if there was not entered data for a tv show. The other columns that we cleaned the values for were age_certification, imdb_score, and imdb_values, and these were cleaned through using the mode value for age_certification and entering "No Data" for the other two columns. For the other columns with null values, their percentages seemed almost negligible for our investigation's purposes, so they remained unaltered.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data %*
id	object	*0.00%*
title	object	*0.02%*
type	object	*0.00%*
description	object	*0.31%*
release_year	int64	*0.00%*
age_certification	object	44.77%
runtime	int64	0.00%
genres	object	0.00%
production_countries	object	0.00%
seasons	float64	64.00%
imdb_id	object	6.89%
imdb_score	float64	8.24%
imdb_votes	float64	8.51%
tmdb_popularity	float64	1.56%
tmdb_score	float64	5.32%

III. Data Set Summary Statistics

Below are statistical summaries from our selected dataset. In Table 2, statistical measurements measuring centrality, minimum and maximum values, standard deviation, and values of the first and third quartiles. These measurements aid in findings regarding any possible skew there might be in the data, as well as help to showcase any anomalies in the dataset. Table 3 displays the proportion and frequency of the different age certifications the titles on Netflix as of July 2022 were rated. They are shown in descending order to also aid in showing the skewed distribution towards TV-MA age certifications for programs on the streaming service. Table 4 highlights the distribution of the titles between the types of movie and show. This table also is formatted to show frequency and proportion of the total amount of titles on the platform. The heatmap and corresponding correlation matrix highlight just how little correlation appears to be present between different variables within the dataset. The greatest positive correlation is only 0.043, which is not an extraordinarily strong correlation, with others being so close to zero they could be said to be equal to zero Pearson correlation at all.

Table 2: Summary Statistics for Netflix TV Shows and Movies.

	release_year	runtime	seasons	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
count	5850	5850	2106	5368	5.35E+03	5759	5539
mean	2016.417094	76.888889	2.162868	6.510861	2.34E+04	22.637925	6.829175
std	6.937726	39.002509	2.689041	1.163826	9.58E+04	81.680263	1.170391
min	1945	0	1	1.5	5.00E+00	0.009442	0.5
25%	2016	44	1	5.8	5.17E+02	2.7285	6.1
50%	2018	83	1	6.6	2.23E+03	6.821	6.9
75%	2020	104	2	7.3	9.49E+03	16.59	7.5375
Max	2022	240	42	9.6	2.29E+06	2274.044	10

Table 3: Proportions for age_certification

Category	*Frequency*	*Proportion (%)*
TV-MA	*3502*	*59.68%*
R	*556*	*9.50%*
TV-14	*474*	*8.10%*
PG-13	*451*	*7.71%*
PG	*233*	*3.98%*
TV-PG	*188*	*3.21%*
G	*124*	*2.12%*
TV-Y7	*120*	*2.05%*
TV-Y	*107*	*1.83%*
TV-G	*79*	*1.35%*
NC-17	*16*	*0.27%*

Table 4: Proportions for type

Category	*Frequency*	*Proportion (%)*
MOVIE	*3744*	*64.0%*
SHOW	*2106*	*36.0%*

Heat map and correlation matrix

	release_year	runtime	tmdb_popularity	tmdb_score
release_year	1.000000	-0.173551	0.043244	0.032073
runtime	-0.173551	1.000000	0.002545	-0.277454
tmdb_popularity	0.043244	0.002545	1.000000	0.071894
tmdb_score	0.032073	-0.277454	0.071894	1.000000

IV. DATA SET GRAPHICAL EXPLORATION

In our exploratory analysis, we developed and created multiple charts and graphs highlighting various aspects of the data. The usage of graphical representations in analysis work allows for information to be more easily understood by those who examine the data and the findings. Graphical visualization additionally allows for multiple variables to be compared in a more compact way than through using multiple tables. Below are our findings from our exploration, with visuals coming from Python libraries such as seaborn, pandas, and matplotlib, and other software such as Tableau. All visuals and graphics come after data cleaning waws performed on the dataset.

A. Distributions

Through our exploration, we were able to find the distribution of titles of content on Netflix categorized into two types: movie and show. This can be seen in Figure 1, which will also be discussed later. Additionally, the titles found in the dataset have a wide range of age certification rating, imdb scores and votes, and tmdb scores and popularity ratings. Age certification ratings can be seen represented in Figure 2. Imdb scores and votes, and Tmdb scores and popularity ratings range from 0.0 to 10 as shown above in Table 2. We found that Imdb scores and Tmdb score are not necessarily related directly as those categories have different statistical measurement findings which are also represented in an above section.

B. Scatterplots / Pairwise Plots (continuous variables)

Our exploration allowed us to further investigate the possibility of correlation between variables using scatterplots. In figures 3, 4, and 5 below, scatterplots of seasons versus different continuous variables can be seen. These plots show insight into how different ratings and scores given to titles can be related to something such as how many seasons the show runs for. Pair plots, very much like scatterplots, help to visualize possible correlations between continuous variables but differ by showcasing all possible plots for the specified dataset's continuous variables all within one localized plot. An example of such can be seen in Figure 9 below. The pair plot that we created is further broken down into being color coated by if the titles were movies or television shows.

C. BarCharts (categorical variables)

Barcharts are a graphical representation that not only aids in the further understanding and visualization of distribution, but also frequency of variables. Figure 2 showcases a bar plot with age certification on the x axis, and

the count of titles for that specific certification on the y axis. Categorical variables can be analyzed and understood through such visual representations.

D. Other Plots

A sampling of other plots created by our analyzation of the dataset can be seen below (see Figure 1, Figure 3, and Figure 7). Our usage of a pie chart helps to display the proportion of tv shows more easily to movies in our dataset. Histograms, much like bar charts, help to visualize distribution of data to be more regularly understandable by anyone who looks at the chart. In our usage, we were able to distinguish where the majority of our data was clustered in regard to number of seasons and tmdb scores for example.

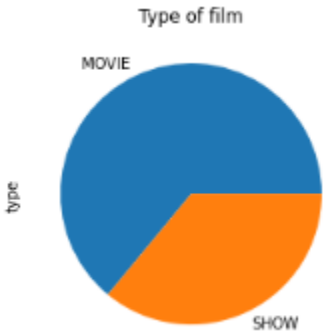

Figure 1: Pie plot for type variable

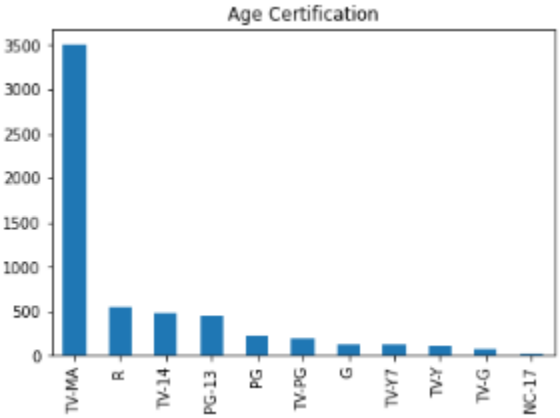

Figure 2: Bar plot for age_certification variable

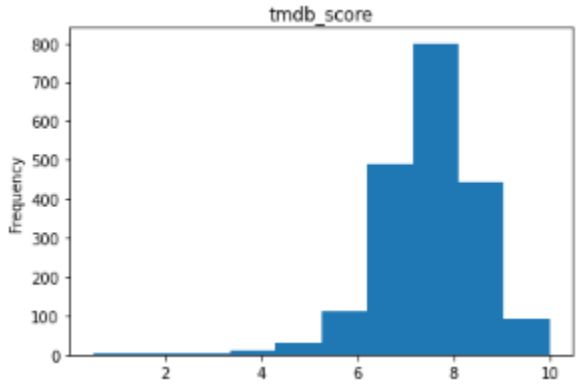

Figure 3: Histogram for tmdb_score variable

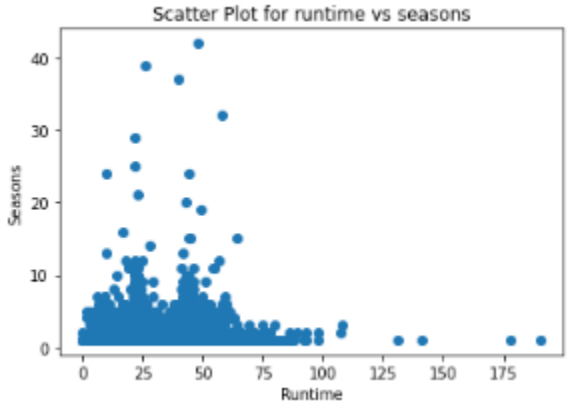

Figure 3: Scatter plot for runtime vs seasons

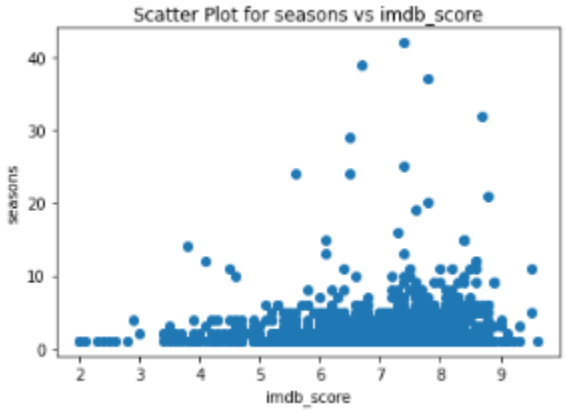

Figure 5: scatterplot for seasons vs imdb_score

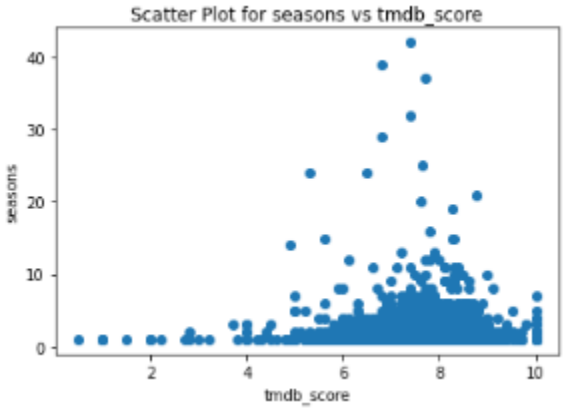

Figure 6: scatterplot for seasons vs imdb_score

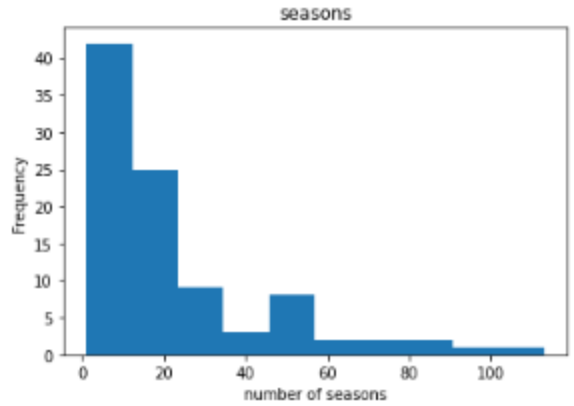

Figure 7: histogram for seasons variable

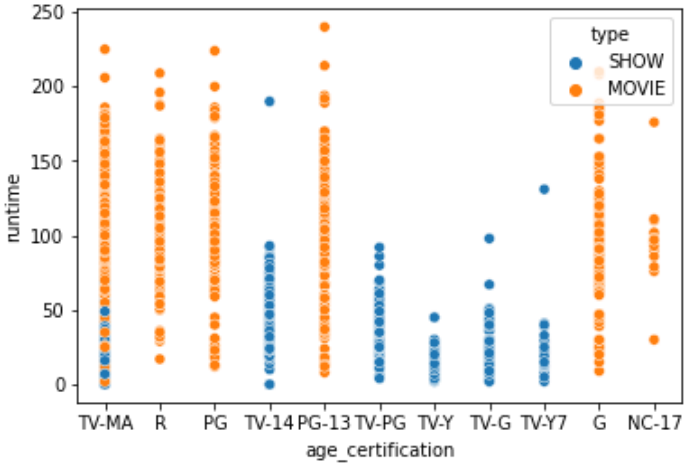

Figure 8: scatterplot for runtime versus age certification

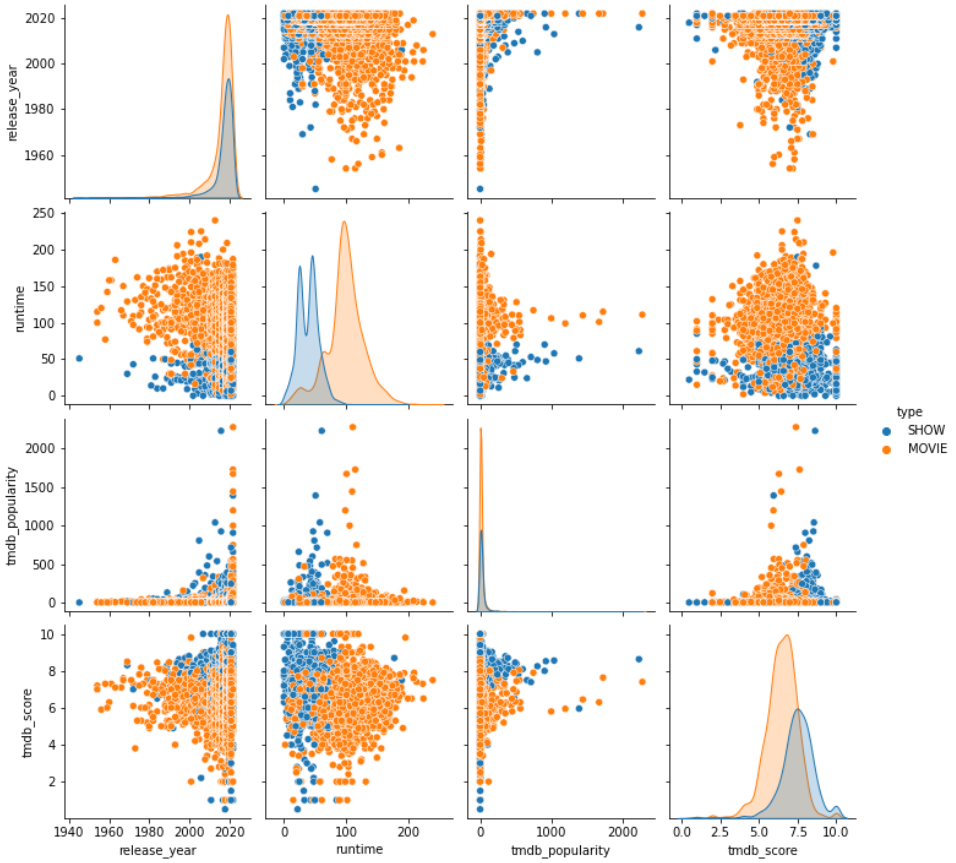

Figure 9: pair plot for dataset, hue='type'

V. SUMMARY OF FINDINGS

We used many different exploratory data analysis techniques to try and find trends and understand our data set. After we cleaned our data, we used multiple libraries to make graphs to display our data. To start, we found that most movies and shows that are currently on Netflix were released between 2010-2020. After that we made a heatmap. In terms of correlation between variables, not much was found. There was a small correlation between tmbp_popularity and tmdb_score, release year and tmdb_popularity, and release year and tmdb_score. By using pandas, we found statistical values for our numerical columns. These values include mean, mode, median, standard deviation, etc. As shown in multiple ways, most content on Netflix comes in the form of movies. Also, the most frequent age certification is TV-MA. We found that tmdb_scores on Netflix are left skewed, with the median being 6.9. Through a scatterplot, you could see that the shows that had multiple seasons were shorter in runtime. Another scatterplot also revealed that shows with a high number of seasons tended to have a higher imdb score. The last scatterplot showed that only movies and shows to receive a ten tmbd_score were less than 100 minutes (about 1 and a half hours) in runtime. Overall, there were not super strong correlations between the data, but we certainly found trends that showed some relationships between different variables.