Exploring Streaming Platforms

Kelsey Kua, Kenneth Ven, Huy Tran

Overview

Online video streaming is commonly used by people of all ages. One of the first big streaming platforms was Netflix, and after their success, other new streaming services popped up. Now there are several services to choose from, and it is not uncommon for an individual consumer to subscribe to multiple streaming services.

In this project, we will be exploring the availability of movies on popular streaming platforms such as Netflix, Prime Video, and Hulu.

Area of Application

Our area of application is about analyzing different streaming platforms to see which movies platforms prefer to seek out. Other streaming platforms or consumers that seek to understand the movies on other platforms may find a use for this project. This project looks at the type of movies currently on the platform from their genres, language, directors, or age group. With this analysis, other streaming platforms and consumers can see which platform they or others would prefer.

Our Approach

The question we want to answer is the following: **What kind of content does each major streaming platform tend to feature (e.g. genre, language, age group, year)?** To go about answering this question, we will be using public datasets found online. The current datasets that we will use can be found here from Kaggle (last updated: 5 months ago). There has been a report that can be found here that describes the accuracy and intent of the current data.

For future and additional datasets, we will find non-noise sources online that match the requirements for what we are testing and then download them to use in our project.

From this dataset, we will examine the following variables:

1. Movie titles
2. Movie year
3. Target age group (per movie)
4. Genre
5. Platform
    a. Netflix
    b. Hulu

    c. Disney+
    d. Prime Video
6. Movie Rating
    a. IMDb
    b. Rotten Tomatoes
7. Language
8. Director

With these variables, we hope to identify the content patterns within each platform to provide a high-level summary of media content. The features we can control include the streaming platform, genre, movie year, target age group, and language. For example, we can see how the variety of genre, movie year, age group, and language differ by platform. The feature we cannot control is ratings (IMDb and Rotten Tomatoes), which have been determined by users of these websites.

Measurement of Success

We want to be able to create a way for consumers to easily see which streaming platform they would prefer to use as well as help compare platforms to see which movies they do not have. The data collection we want to achieve would create an accurate view of streaming from examining a large spread of movies within each platform. From these measurements, we can be able to determine the types of movies each platform prefers to help read which movies a platform might want to buy next.

Assumptions

We are assuming that the runtime listed in the dataset is the runtime of the most common version of the film. For example, every Harry Potter movie (let's say the first one) that has been watched by people is the same, so no deleted scenes or some factor that might add time to a version that makes it a different runtime from what someone might watch.

We are assuming that a movie rating reflects viewers' overall opinions of the movie. That is, the higher the rating, the better the movie.

We are assuming that the target age group is determined by the film's official rating. For example, a rating of PG-13 yields the value of '13+' in the dataset.

We are assuming that all the movies in the dataset are being rated on the same scale (e.g 1-5 star rating)

Potential Problems With the Data (noise sources)

If using the Kaggle dataset, some noise sources will occur. One of these sources is that some data may be inaccurate (e.g. runtime). This noise source might affect what variables we can use if most data of a variable is inaccurate. We also may be unaware that a variable such as runtime is inaccurate; making the data more accurate would require manual checking, which is time-consuming and not feasible to do.

Another noise source is that there may be many null values for a given variable. Upon inspection, we saw that the age group and Rotten Tomatoes ratings have mostly null values. Having missing data would yield inaccurate analysis and make some variables unusable for analysis.

Another noise source is that the streaming platforms' content changes frequently. The data in the Kaggle dataset was only accurate at the moment it was scrapped. This would affect the accuracy of our final results if a platform has added or no longer has a movie from the time the data was scraped.

Team Assessment

Kelsey - I have attended team meetings. I have contributed to our iterations by thinking through our approach, weighing its feasibility, and refining it accordingly. I looked through the kinds of data available in our chosen dataset to determine control variables and noise sources. After becoming familiar with the dataset, I came up with assumptions.

Kenneth - I showed up for the team meetings and provided my contributions to the iterations. When the team finished our work, I revised the document by skimming it over to make sure we covered everything. I looked over the datasets my teammates have found to make sure they are usable and aren't noise sources.

Huy - I showed up to the team meetings on time and contributed to providing information in the iteration. I help collect the data and research the accuracy in the data. I provided ideas for each question when we were refining the technical approach to our project. I skim through the document to discuss anything questionable or did not seem realistic in the project.