Project Description
MATH 7360 – Fall 2021
September 22, 2021

Abigail M Zion

September 22, 2021

1 Introduction

We will be using the dataset **TV Shows and Movies listed on Netflix**.

1.1 Origin

TV Shows and Movies listed on Netflix was retrieved from Shivam Bansal on www.kaggle.com (https://www.kaggle.com/shivamb/netflix-shows). The data is collected from the cite Flixable (https://flixable.com), a search engine for video streaming services. Bansal's dataset consists of 7,787 movies and television shows on Netflix. It was created on December 4th, 2019, and was last updated on January 18th, 2021.

1.2 Data Type

The data has 12 columns, or variables: 10 String, 1 DateTime, and 1 Integer.

1.2.1 String

The String variables are:

1. show_id: There are 7787 unique identifiers; one for every movie or television show.
2. type: Every observation is identified to be either a Movie or TV show.
3. title: Each Movie or TV show has a unique title.
4. director: The director of the Movie. TV shows are [null] in this column.
5. cast: A string listing actors involved in the Movie or TV show.

6. country: The country where the Movie or TV show was produced.
7. rating: The TV rating of the Movie or TV show. For example,"TV-MA".
8. duration: For movies, the duration in minutes is reported. For TV shows, the number of seasons is reported.
9. listed_in: The genre of the Movie or TV show. For example, "Documentaries".
10. description: The summary description of the Movie or TV show.

1.2.2 DateTime

The DateTime variable is:

1. date_added: The date the Movie or TV show was added to Netflix, in the format dd-Mon-yy. For example, "14-Aug-20". The dates range from December 31, 2007 to January 15, 2021.

1.2.3 Integer

The Integer variable is:

1. release_year: The year that Movie or TV show was actually released. The years range from 1925-2021.

2 How We Chose This Dataset

We chose this dataset based on it's usability, size, and availability of questions to be asked.

The dataset contains a managable number of variables. We find that all variables listed are relevant and informative.

With 7,787 observations, this dataset is large enough that it cannot be understood without statistical analyses, yet small enough to avoid computational lags.

Most of the variables are predictor variables, while we have one key outcome variable: the length of time between the release date and the date the Movie or TV show was added to Netflix. Many of the datasets we considered did not have both predictors and outcomes, which are useful for statistical analyses. Additionally, the description of the Movie or TV show can be analyzed against other variables such as the genre. Finally, the binary variable "type" can be used to make comparisons between Movies and TV shows.

3 Questions

We separate questions based on the their level of complexity and application of statistical analyses.

3.1 Basic

Basic questions can be answered using elementary filtering and sorting techniques. Our basic questions are:

1. Which director has the most movies on Netflix?
2. Which actors are in the most movies? Which actors are in the most TV shows?

3.2 Intermediate

Intermediate questions require knowledge of statistical analyses, perhaps already known to the author. Our intermediate questions are:

1. How do movie genres vary between Movies and TV shows?
2. How has the length of time between a Movie or TV's release date and it's addition to Netflix varied over time?

3.3 Advanced

Advanced questions are more challenging and require data analysis techniques learned in MATH-7360 with Professor Xiang Ji. Our advanced questions are:

1. What key descriptor words are indicative of a Movie or TV shows's genre?
2. What key descriptor words are indicative of a Movie or TV's TV rating?
3. What variables, if any, are indicative of the length of time between a Movie or TV's release date and it's addition to Netflix?