IST 719: Work in Progress Report

Shashank Nagaraja

Data

This dataset contains information roughly ~7,000 titles that are available on the video streaming service Netflix. This represents all titles that Netflix currently had/planned to have as of 2019. The dataset contains additional information about the movie/tv show including the following:

- **Type**: Movie/TV Show
- **Title**
- **Director**: Director(s) of the movie
- **Cast**: Actors in the movie
- **Country**: Country of production
- **Date Added**: Date it was added to Netflix
- **Release Year**: Release year of the production
- **TV Rating**: MPAA Rating
- **Duration**: Total duration in minutes or number of seasons

Data source: https://www.kaggle.com/shivamb/netflix-shows

Compelling Story

With an ever-growing repertoire of movies and TV Shows, Netflix has grown to become one of the largest video streaming platforms in the world. However, it is interesting to note that Netflix (instead of simply adding more and more movies) strategically balances out their offerings based on viewing habits and user requests.

Audience

Avid movie watchers and critics will find the findings of this analysis interesting. Since the service is now part of most people's daily lives, it is representative of viewing habits in general.

Exploration

In order to explore the data, it was first cleaned. Although most fields contained the data in a string format, some of the fields had to be parsed from their date and comma-separated list formats. While a DataFrame object works great for flat data, it was not suitable for the Netflix data that was more complex; therefore, the data was stored in a dictionary format that allowed parsing of lists. The data transformation steps were as follows:

1. Data read in as a DataFrame

2. Column ‘show_id’ dropped as it serves no purpose
3. Various metrics computed using columns that are not lists
4. Iterated over DataFrame and created dictionary for each movie
5. Fixed list columns
6. Computed multiple contingency tables for further plotting
7. Plotting

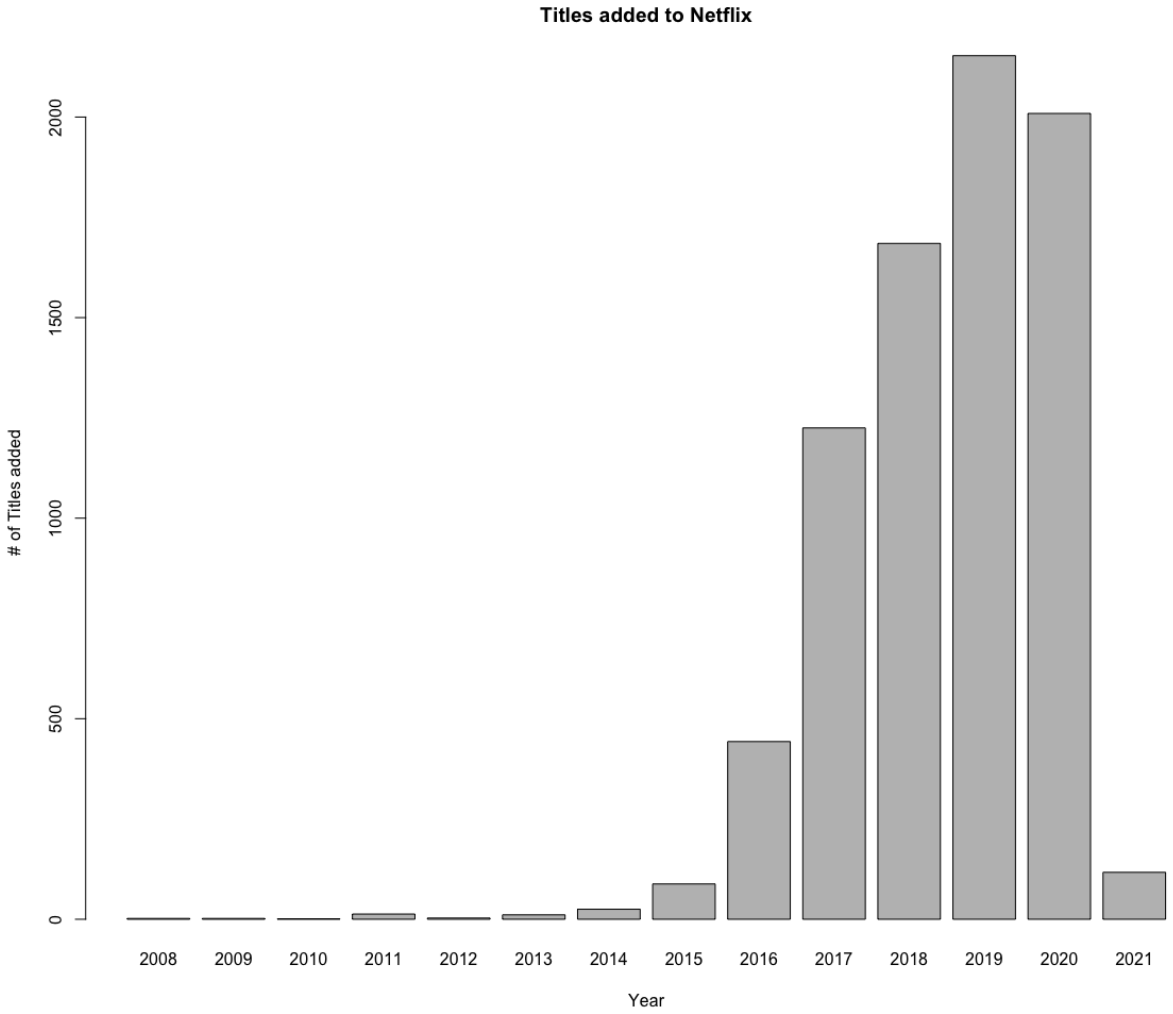
Questions

1. **How many titles has Netflix added per year?**
   Unit of Analysis: Production (Movie/TV Show)
   Comparison Values: Year Added to Netflix
   Computed with: Count
2. **Does Netflix add more TV Shows, or Movies?**
   Unit of Analysis: Production (Movie/TV Show)
   Comparison Values: Production Type
   Computed with: Count
3. **What is the distribution of Genres?**
   Unit of Analysis: Genre
   Comparison Values: Year Released
   Computed with: Count
4. **What is the distribution of the country of Production?**
   Unit of Analysis: Country of Production
   Comparison Values: Year Released
   Computed with: Count

Plots

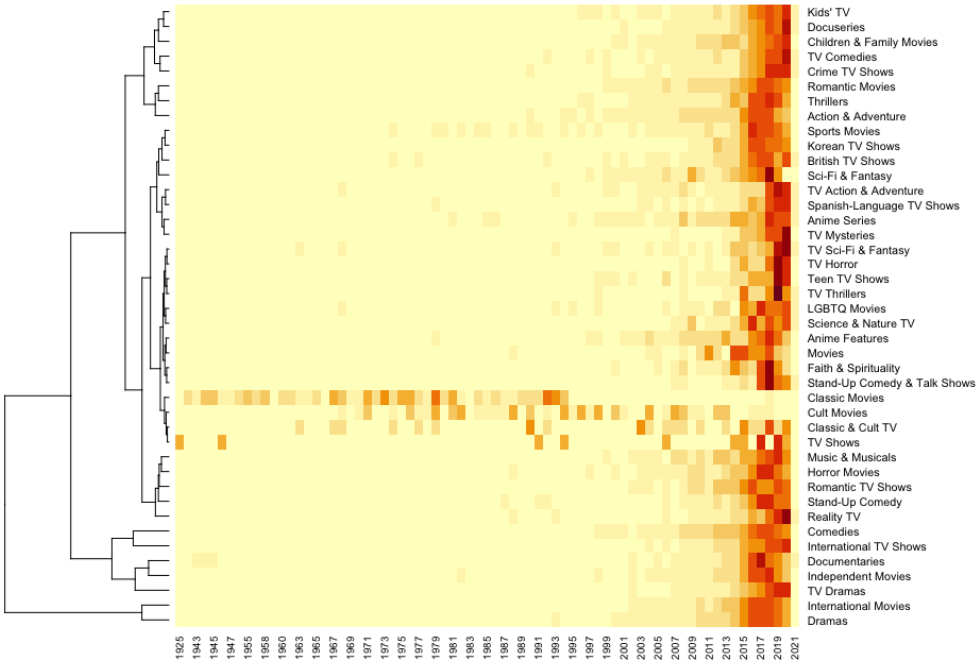
My program (main.R) contains functions for the steps described in the ‘Exploration’ section of this report. This method of neatly packaging code into reusable functions that can be strung together is one of the advantages of working with a language like R. The main() function calls these various functions and passes along the data, transforming the data along the way. The final output files are written as CSV files. They are:

1. **How many titles has Netflix added per year?**
   date_added_toNetflix.csv
   Table of total number of productions added to Netflix per year.



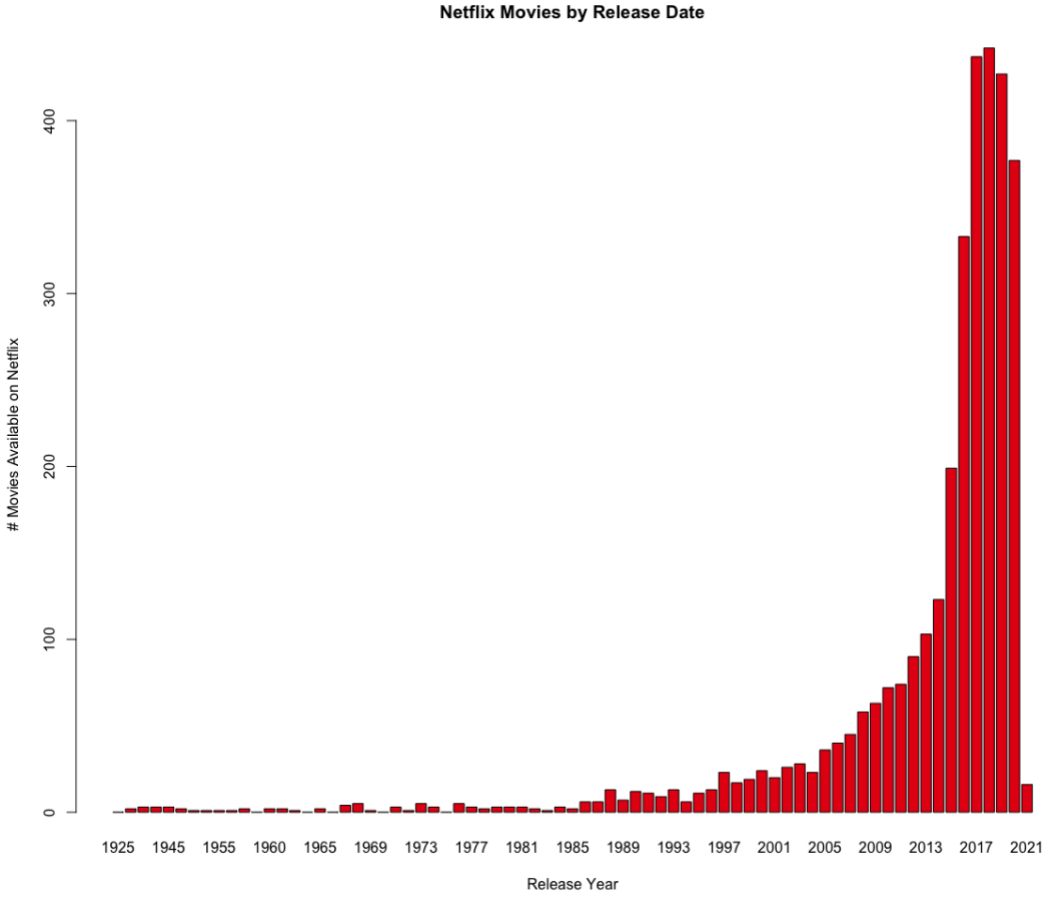
2. **Does Netflix add more TV Shows, or Movies?**
   Printed output.
   Netflix has about twice as many Movies(5377) as TV Shows(2410).

3. **What is the distribution of Genres?**
   listed_in_vs_release_year.csv, GenreYear.png
   Interestingly, there has been a decline in Action & Adventure, and SciFi Movies to give way to more TV Mysteries and Reality TV on Netflix.



4. **What is the distribution of the country of Production?**
   Country_vs_release_year.csv
   *Unsurprisingly, most movies available on Netflix are produced in the United States. India is also a large producer of productions, probably due to its large Bollywood industry. United Kingdom follows these two countries in production numbers. *

Additional Plots:


Netflix Movies by Release Date
Movies Available on Netflix
0
100
200
300
400
1925
1945
1955
1960
1965
1969
1973
1977
1981
1985
1989
1993
1997
2001
2005
2009
2013
2017
2021
Release Year

***