Data Analysis Project Part I: Netflix Movies and TV Shows Exploratory Analysis

Is It Over Yet?
How Factors Affect
Movie/Series Length

Applied Statistical Computing Fall 2022
Professor Anthony Donoghue

Table of Contents

I. Introduction

The dataset used in this project is the Netflix Movies and TV Shows dataset created by Shivam Bansal and shared through Kaggle. This dataset contains over 8800 listings of the movies and series on Netflix and their corresponding title, director, country of production, date added to netflix, original release date, TV rating, genre, description and runtime. The dataset was last updated September 25, 2021 and contains listings up until that date. In this project, I would like to determine the factors that are most correlated to runtime by exploring runtime in reference to different qualitative factors including the directors, countries of production, TV ratings, genre and description. Then I will further explore more possible correlations by observing runtime in reference to the quantitative variables including the release year and date added to Netflix.

After this initial analysis, I will observe the factors that appeared to have strongest correlation in reference to each other in hopes of gaining a better understanding to why these factors correlate and develop hypothesis' for future statistical testing.

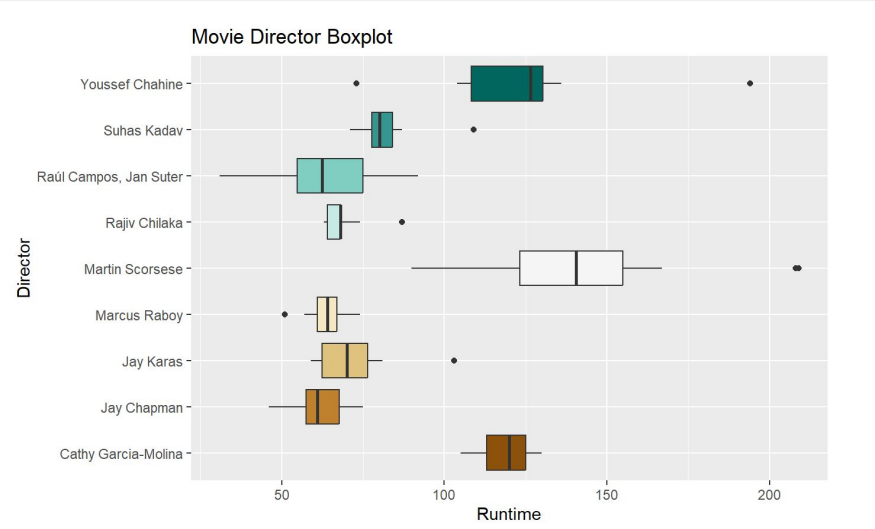
The purpose of this analysis is to gain a better understanding of why some movies and series have much longer runtimes than others and identify factors to look for when searching for something new to watch to match showtime length preferences.

II. Qualitative Variables

II. A. Director

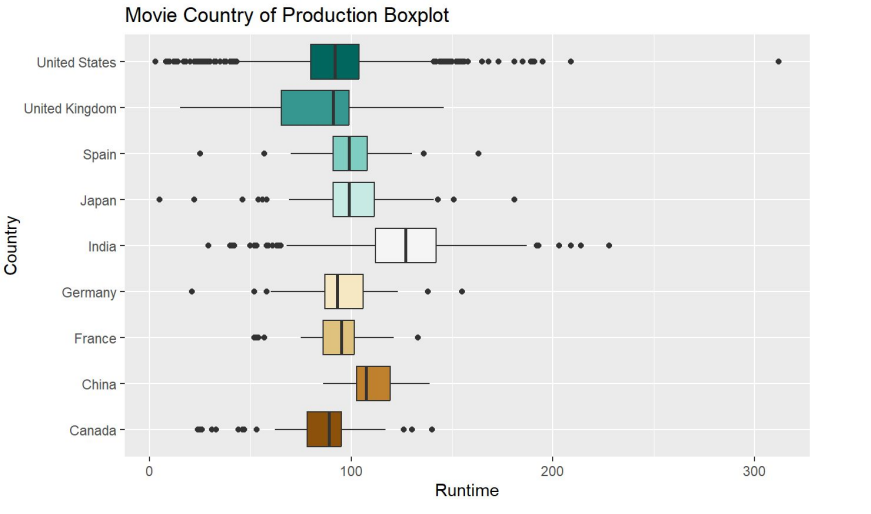
Perhaps different directors tend to make films of a certain length. Here I used the data to find the directors with the most amount of films in the dataset and analyzed them in regards to runtime measured in minutes. There appears to be a strong correlation between these variables as directors like Martin Scorsese (144 min on average) tend to make much longer films than directors like Jay Chapman (62 min on average). However The number of films which this data is based on is also quite low as shown in the table containing the director names and their corresponding number of films. These findings could likely benefit from further exploration to determine if these differences are in fact statistically significant. Unfortunately, I could not do this same analysis for TV show directors as even the directors with the greatest amount of works only had at most 3 TV shows.

Rajiv Chilaka	19
Raúl Campos, Jan Suter	18
Suhas Kadav	16
Marcus Raboy	15
Jay Karas	14
Cathy Garcia-Molina	13
Jay Chapman	12
Martin Scorsese	12
Youssef Chahine	12




```
$`Cathy Garcia-Molina`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  105.0   113.0   120.0   118.2   125.0   130.0

$`Jay Chapman`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   57.50   61.00   61.67   67.75   75.00

$`Jay Karas`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.00   62.50   70.00   71.14   76.50  103.00

$`Marcus Raboy`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   51.0    61.0    64.0    63.8    67.0    74.0

$`Martin Scorsese`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   90.0   123.2   140.5   144.2   155.0   209.0

$`Rajiv Chilaka`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  63.00   64.00   68.00   67.84   68.00   87.00

$`Raúl Campos, Jan Suter`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.00   54.75   62.50   63.61   75.00   92.00


$`Suhas Kadav`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   77.75   80.00   81.69   84.00  109.00

$`Youssef Chahine`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   73.0   108.2   126.5   123.5   130.2   194.0
```

II. B. 1. Country of Production: Movies



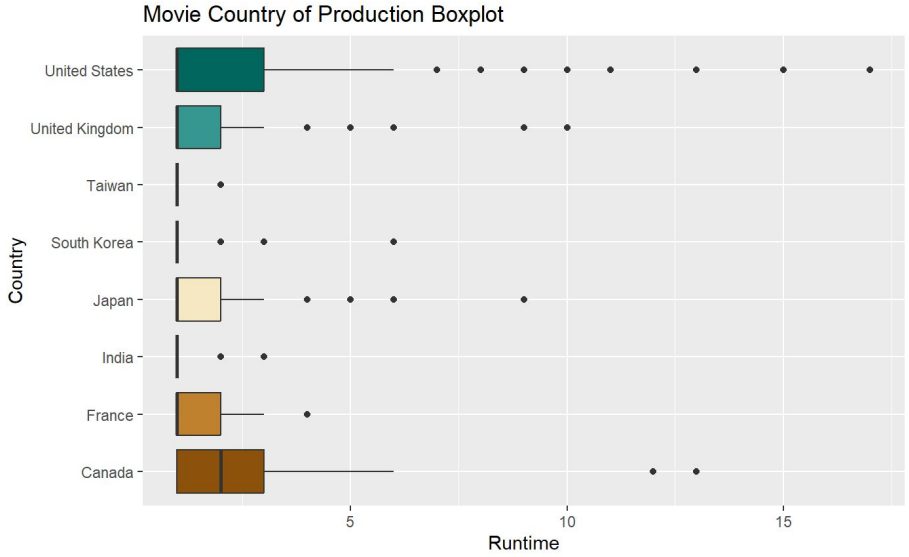
United States	2751
India	962
United Kingdom	532
	441
Canada	319
France	303
Germany	182
Spain	171
Japan	119
China	114

Perhaps different cultures influence the duration of a film. Here I used the data to find the countries with the most amount of films in the dataset to analyze how the country of production may influence the duration of a movie. While the correlation does not seem as prevalent as with the directors, with each country having much more movies than each director, the difference does not need to be as large to be statistically significant. And there does appear to be some correlation between the variables as countries like India (127 min average) tend to have much longer movies than countries like Canada (83 min average). Still to determine whether is this enough, further exploration could be very helpful.

```
$Canada
   Min. 1st Qu.  Median    Mean 3rd Qu.
  24.00   78.25   89.00   82.73   95.00
   Max.
 140.00

$China
   Min. 1st Qu.  Median    Mean 3rd Qu.
   86.0   102.8   107.5   109.9   119.2
   Max.
  139.0

$France
   Min. 1st Qu.  Median    Mean 3rd Qu.
  52.00   86.00   95.00   93.47  101.50
   Max.
 133.00

$Germany
   Min. 1st Qu.  Median    Mean 3rd Qu.
  21.00   87.00   93.00   93.98  106.00
   Max.
 155.00

$India
   Min. 1st Qu.  Median    Mean 3rd Qu.
   29.0   112.0   127.0   126.9   142.0
   Max.
  228.0

$Japan
   Min. 1st Qu.  Median    Mean 3rd Qu.
   5.00   91.00   99.00   98.46  111.50
   Max.
 181.00

$Spain
   Min. 1st Qu.  Median    Mean 3rd Qu.
   25.0    91.0    99.0   100.1   108.0
   Max.
  163.0

$`United Kingdom`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  15.00   65.25   91.00   84.87   99.00
   Max.
 146.00

$`United States`
   Min. 1st Qu.  Median    Mean 3rd Qu.
   3.00   80.00   92.00   90.63  104.00
   Max.    NA's
 312.00       3
```

II. B. 2. Country of Production: TV Shows

```
$Canada
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.576   3.000  13.000

$France
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.367   2.000   4.000

$India
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.165   1.000   3.000

$Japan
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    1.00    1.58    2.00    9.00

$`South Korea`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.177   1.000   6.000

$Taiwan
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.015   1.000   2.000

$`United Kingdom`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.901   2.000  10.000

$`United States`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   2.333   3.000  17.000
```

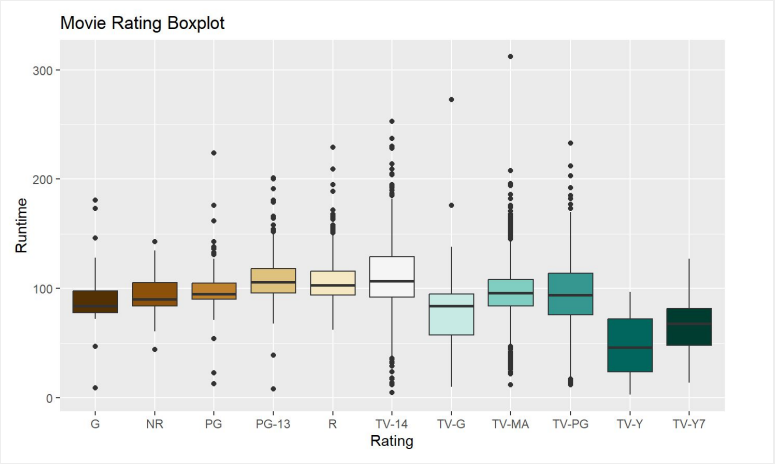


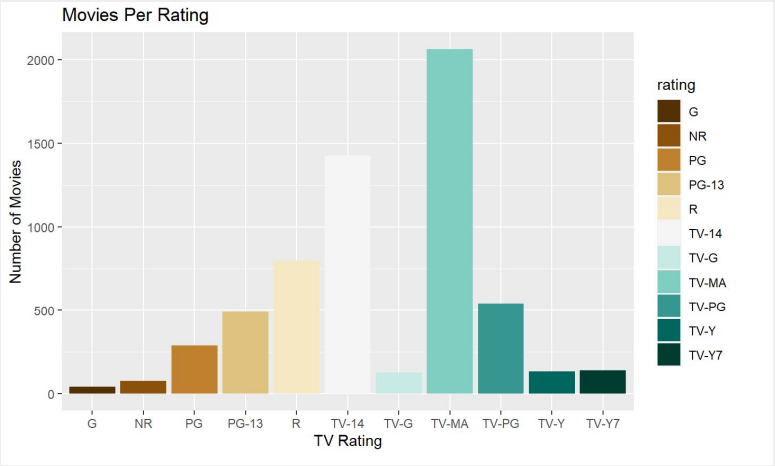
United States	938
	392
United Kingdom	272
Japan	199
South Korea	170
Canada	126
France	90
India	84
Taiwan	70

As with movies, perhaps country of production could have some effect on the length of a series where runtime is measured in seasons of the series. Using the countries with the most TV shows in the dataset, there does somewhat seem to be a slight correlation as countries like Canada (2.5 seasons on average) tend to have significantly longer TV shows than countries like India (1.2 seasons on average). However when looking at the boxplot, this correlation does not appear to be as present as near every country has a 1 season median.

II. C. 1. TV Rating: Movies

```
$G
   Min. 1st Qu.  Median    Mean 3rd Qu.
   9.00   78.00   84.00   90.27   98.00
   Max.
 181.00

$NR
   Min. 1st Qu.  Median    Mean 3rd Qu.
  44.00   84.00   90.00   94.53  105.50
   Max.
 143.00

$PG
   Min. 1st Qu.  Median    Mean 3rd Qu.
  13.00   90.00   95.00   98.28  105.00
   Max.
 224.00

$`PG-13`
   Min. 1st Qu.  Median    Mean 3rd Qu.
    8.0    96.0   106.0   108.3   118.0
   Max.
  201.0

$R
   Min. 1st Qu.  Median    Mean 3rd Qu.
   62.0    94.0   103.0   106.7   116.0
   Max.
$`TV-14`
   Min. 1st Qu.  Median    Mean 3rd Qu.
    5.0    92.0   107.0   110.3   129.0
   Max.
  253.0

$`TV-G`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  10.00   57.25   84.00   79.67   95.00
   Max.
 273.00

$`TV-MA`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  12.00   84.00   96.00   95.89  108.00
   Max.
 312.00

$`TV-PG`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  12.00   75.75   94.00   94.85  114.00
   Max.
 233.00

$`TV-Y`
   Min. 1st Qu.  Median    Mean 3rd Qu.
   3.00   24.00   46.00   48.11   72.00

$`TV-Y7`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  14.00   48.00   68.00   66.29   81.50
   Max.
 127.00
```

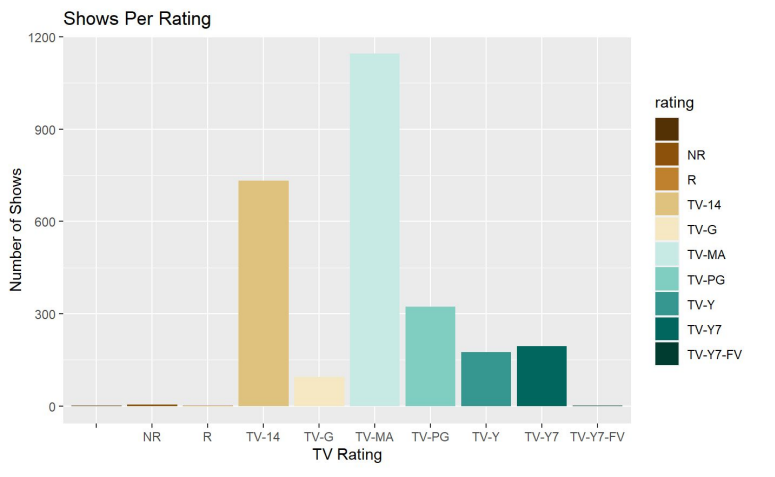


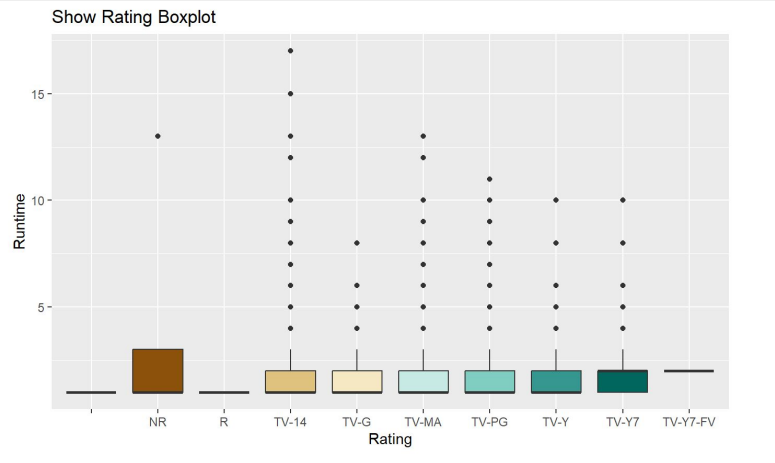


Maybe movies marketed towards different ages/maturities have difference runtimes. Looking at the runtimes for each of the different TV ratings, this could be the case as ratings like TV-14 (114 min average) have much longer runtimes than ratings like TV-Y (48 min average). I would like to look into this data further to determine if this apparent correlation is in fact statistically significant or just a result of the low sampling data of these ratings and if significant, further explore why.

II. C. 2. TV Rating: Shows

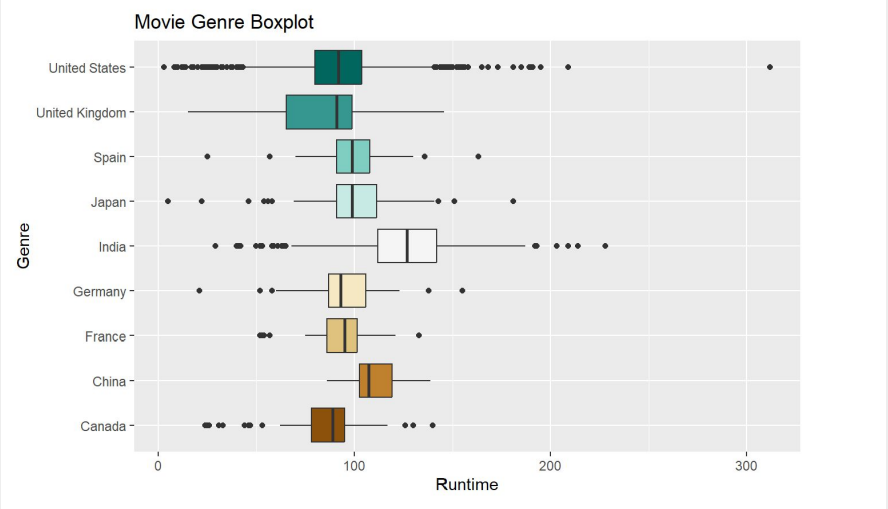
Here I explored ratings in regards to show length. While the means in some ratings do differ slightly (2.0 high end vs 1.6 low end seasons), the boxplot does not show anything particularly interesting beyond a lot of outliers. So there does not appear to be much correlation between tv rating and a shows runtime length measures in seasons.






```
$`TV-14`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  1.000   1.000   1.000   1.821   2.000
   Max.
 17.000

$`TV-G`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  1.000   1.000   1.000   1.851   2.000
   Max.
  8.000

$`TV-MA`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  1.000   1.000   1.000   1.686   2.000
   Max.
 13.000

$`TV-PG`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  1.000   1.000   1.000   1.669   2.000
   Max.
 11.000

$`TV-Y`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  1.000   1.000   1.000   1.852   2.000
   Max.
 10.000

$`TV-Y7`
   Min. 1st Qu.  Median    Mean 3rd Qu.
  1.000   1.000   2.000   2.021   2.000
   Max.
 10.000

$`TV-Y7-FV`
   Min. 1st Qu.  Median    Mean 3rd Qu.
      2       2       2       2       2
   Max.
      2
```

II. D. 1. Genre: Movies

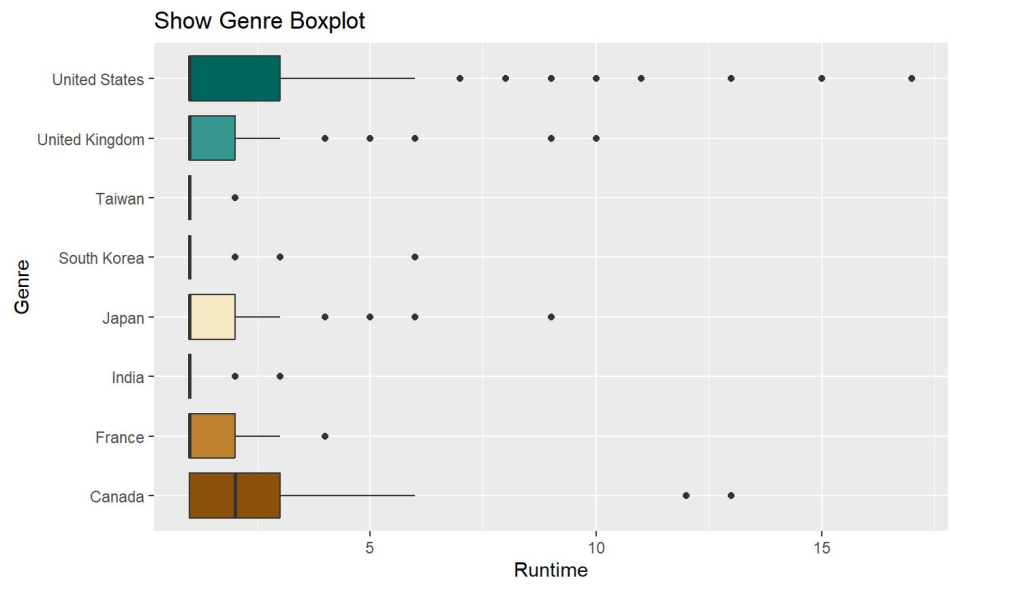


International Movies	2752
Dramas	2427
Comedies	1674
Documentaries	869
Action & Adventure	859
Independent Movies	756
Children & Family Movies	641
Romantic Movies	616
Thrillers	577
Music & Musicals	375

As the correlation between rating and runtime did appear to be somewhat promising for movies, perhaps it may also be interesting to look for some correlation between runtime and genre. Here I looked at the most popular genres and analyzed those. There appears to be a stronger correlation here as genres like Dramas (107 min average) have much longer runtimes than genres like International Movies (61 min average). Further exploration on these variables would likely be beneficial.

```
$`Action & Adventure`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   77.0    94.0   102.0   104.9   113.2   191.0

$`Children & Family Movies`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.00   45.00   68.00   63.56   82.50  139.00

$Comedies
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.00   85.00   93.00   90.23   98.75  190.00

$Documentaries
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   66.50   87.00   80.86   96.00  273.00

$Dramas
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    8.0    95.0   108.5   107.8   121.8   209.0

$`Independent Movies`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    100     100     100     100     100     100

$`International Movies`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50.00   55.50   61.00   61.67   67.50   74.00

$`Music & Musicals`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   52.0    85.5    91.0    94.3   105.0   153.0

$`Romantic Movies`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  81.00   82.50   84.00   83.33   84.50   85.00

$Thrillers
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  80.00   89.00   97.00   99.95  107.00  149.00
```

II. D. 2. Genre: Shows



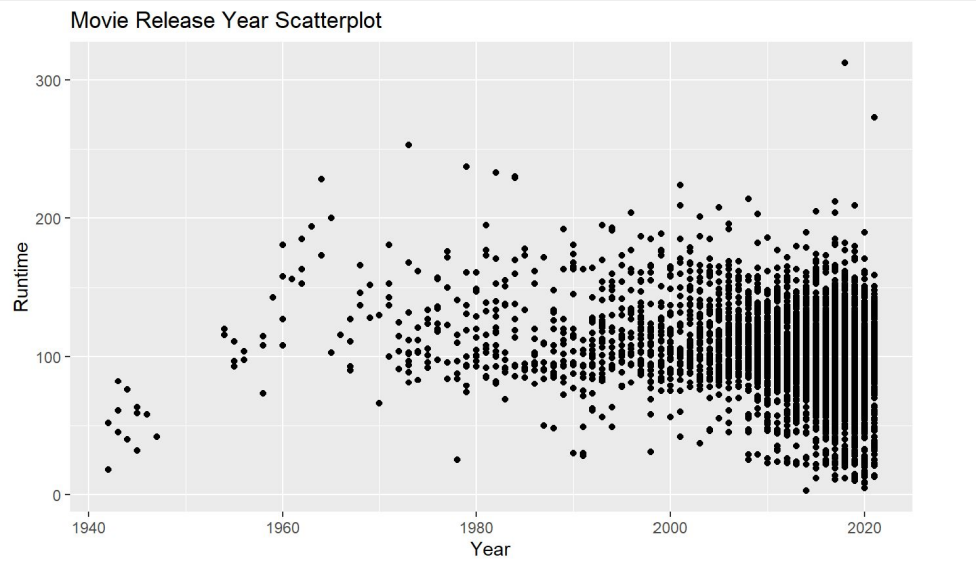

```
$Docuseries
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     1.0     1.0     1.4     1.0     5.0

$`International TV Shows`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       1       1       1

$`Kids' TV`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.882   2.000  10.000

$`Reality TV`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.505   2.000   6.000

$`TV Comedies`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.783   4.000   9.000

$`TV Dramas`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   3.057   4.500  13.000
```

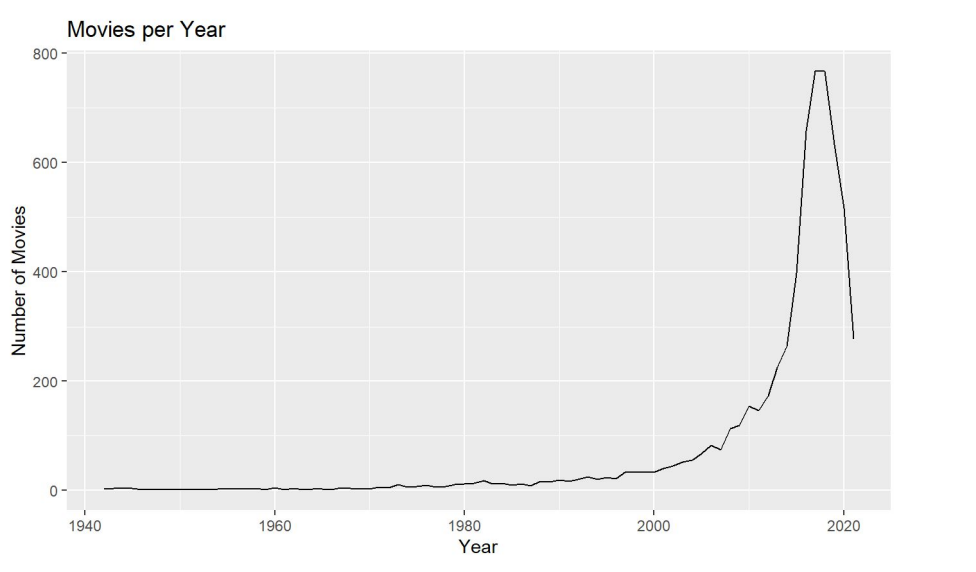
6	International TV Shows	1351
17	TV Dramas	763
16	TV Comedies	581
4	Crime TV Shows	470
7	Kids' TV	451
5	Docuseries	395
10	Romantic TV Shows	370
9	Reality TV	255
2	British TV Shows	253

Despite the lack of much correlation between ratings and TV show runtime, I explored to see if there was any correlation between show runtime and genre. This turned out to be much more correlated as genres such as TV Dramas (3.1 seasons average) are much longer than International TV shows (1 season average). Further exploration using these variables could also be beneficial.

II. Quantitative Variables

III. A. 2. Release Year: Movies

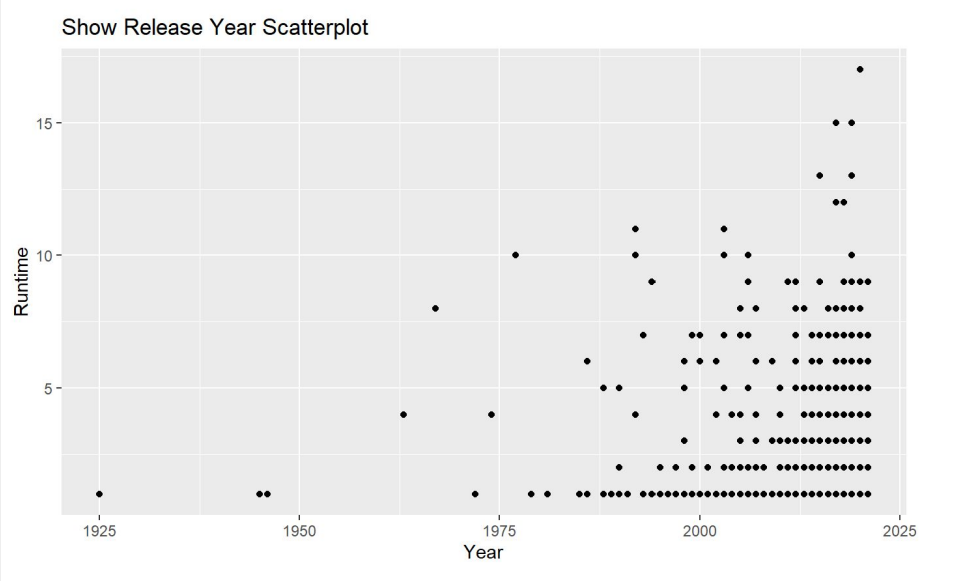


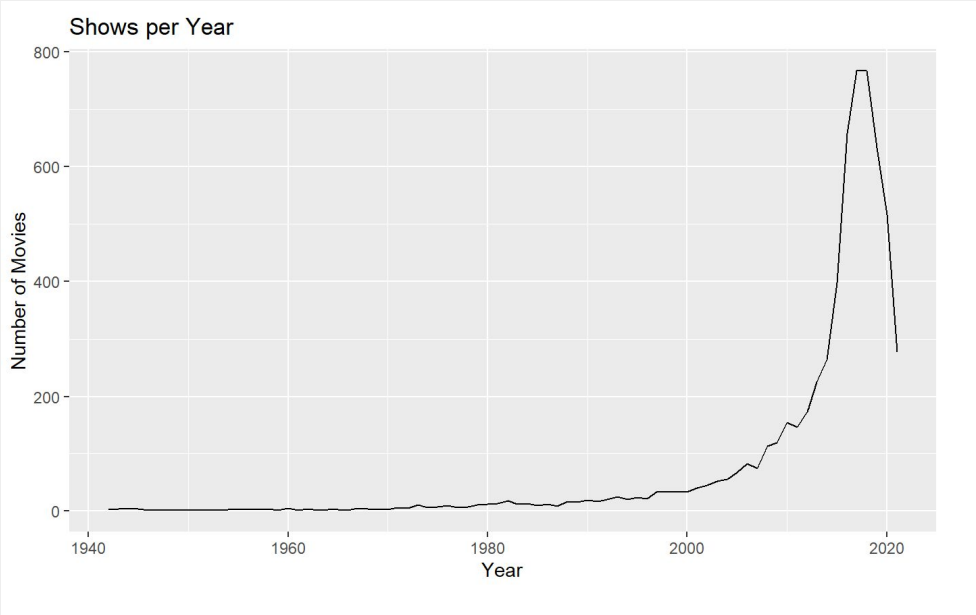


Maybe the length of movies has just increased with time. Here I looked at the runtimes for movies in reference to the year in which they were originally released. Thanks to the massive growth in current movies as seen in the Movies per Year graph, it is a bit difficult to make out if there is a correlation or not. Running further examinations onto this data may be helpful for determining that. Note: The drop in movies around at 2019 may be due to this being a Netflix dataset and Netflix mostly adds movies that have already been release for a while.

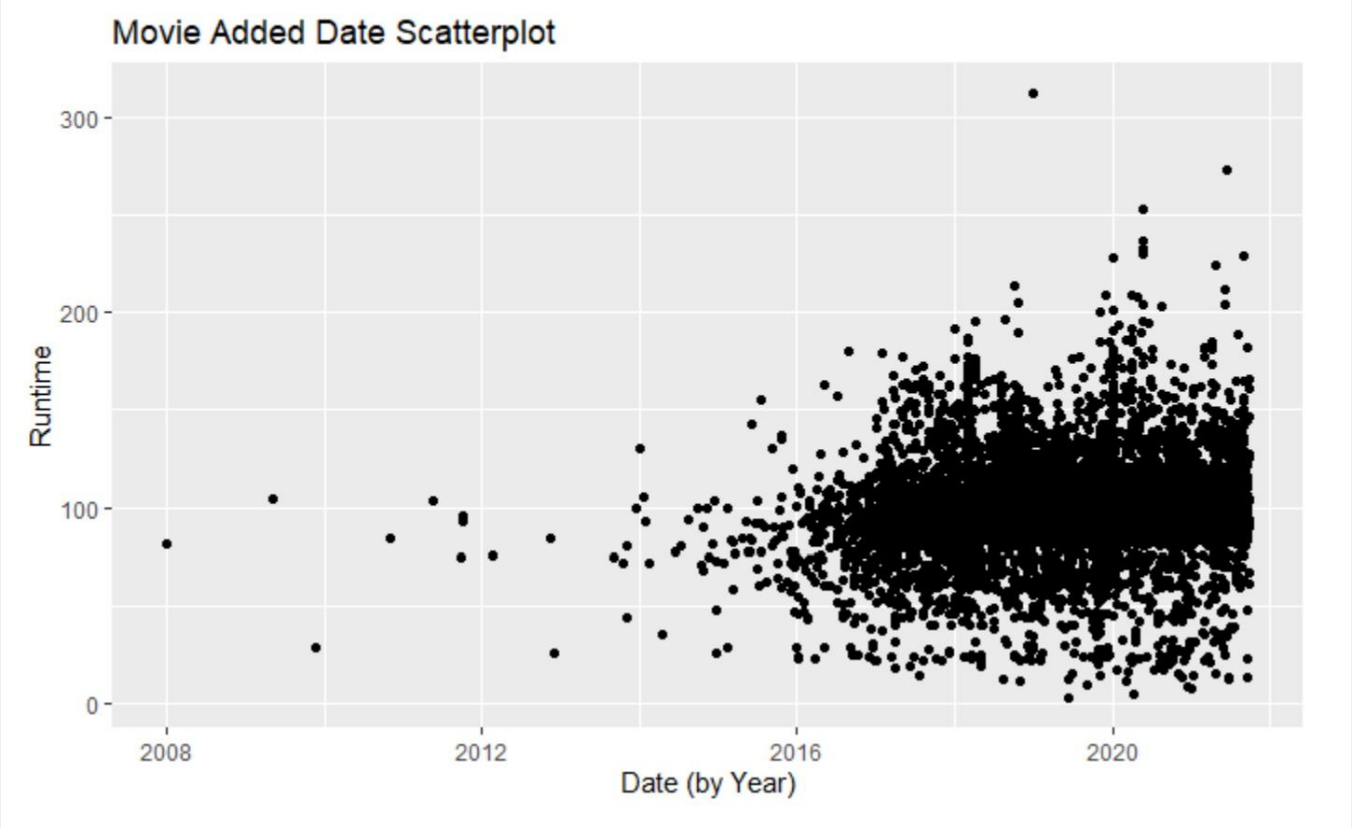
III. A. 2. Release Year: Shows

Okay movies may not be getting longer, but surely TV Shows are, right? Probably not. It's difficult to tell from this scatterplot for the same reasons as before and further investigation would be helpful but at first glance there does not seem to be a correlation. However it should be noted that when regarding shows, current shows may still be in production and thus will eventually have more seasons which is not reflected in the data I used for this analysis.



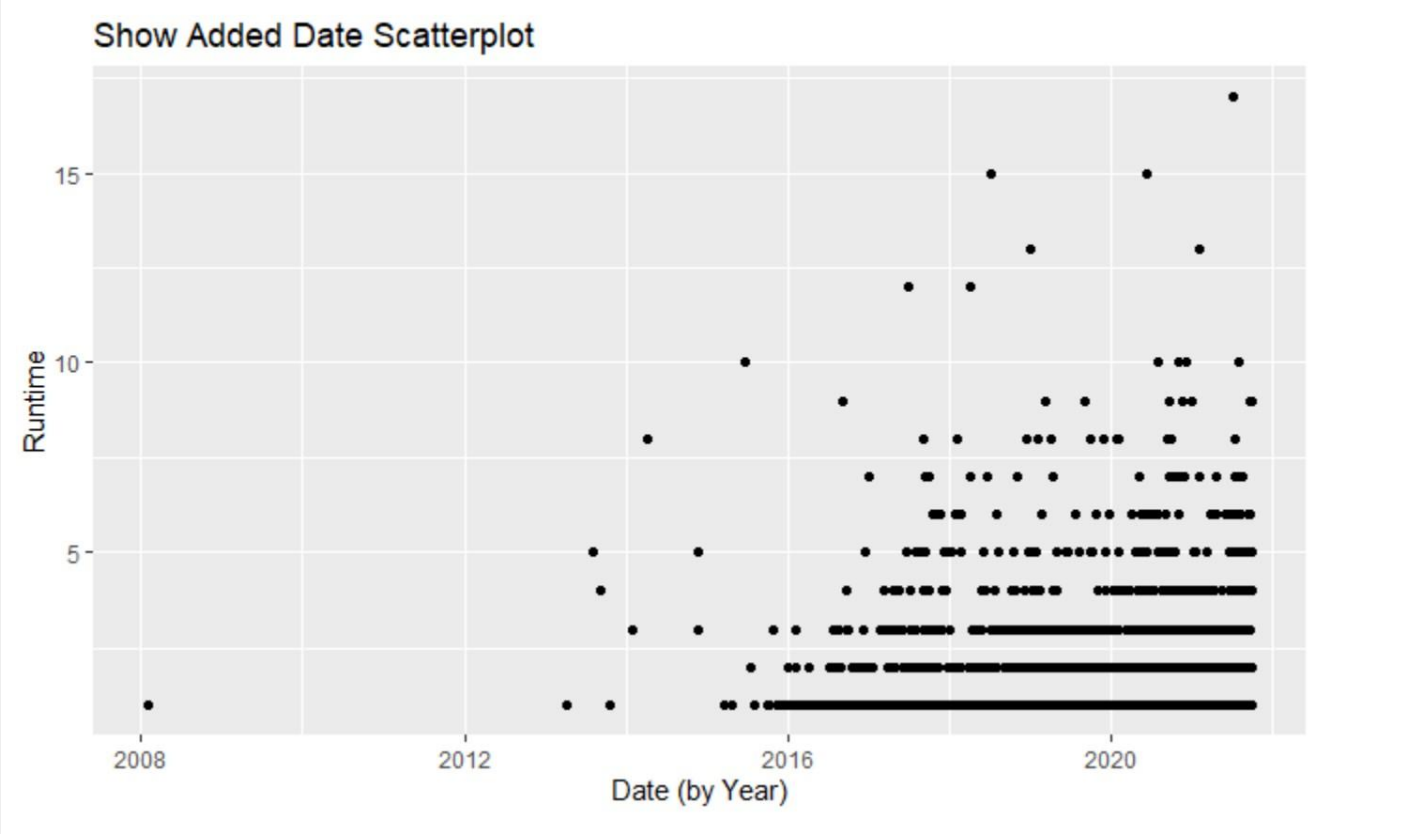


III. B. 1. Date Added to Netflix: Movies



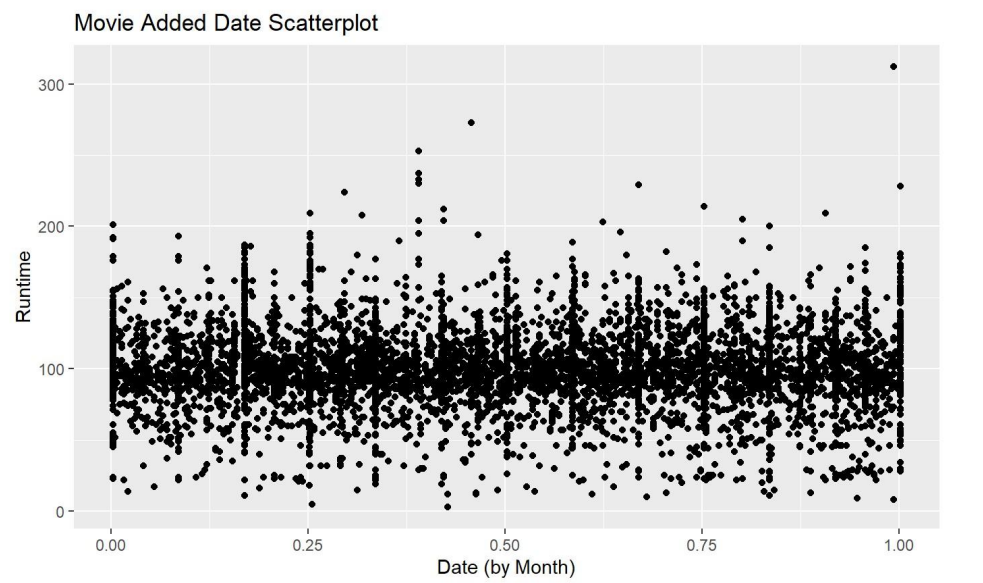
If movies are inherently getting longer, maybe Netflix is just choosing to add longer movies now. This could be the case and if I squint at the scatterplot above, there does appear to maybe be a slight upward trend. With so many movies, perhaps this is enough to be statistically significant but its difficult to say with much certainty at this point. I think this is also something worth exploring further.

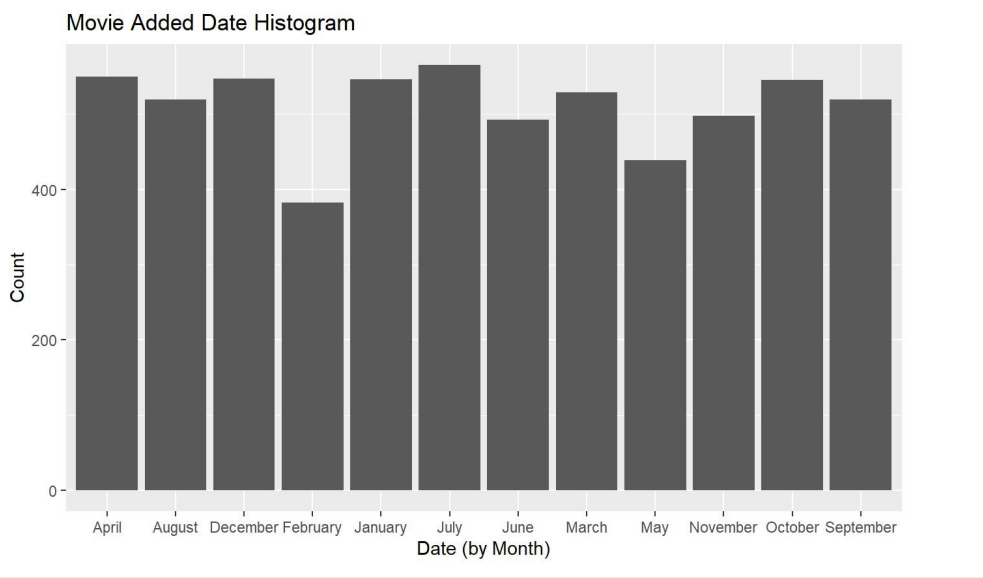
III. B. 2. Date Added to Netflix: Shows



Could it be the case that Netflix is adding/making longer TV shows? It's still difficult to say but there does not appear to be much or any correlation between the date a show was added to Netflix and its corresponding number of seasons. However, like before, it is important to consider that recent shows may have many more seasons in the future.

III. C. 1. Time of Year Added to Netflix: Movies

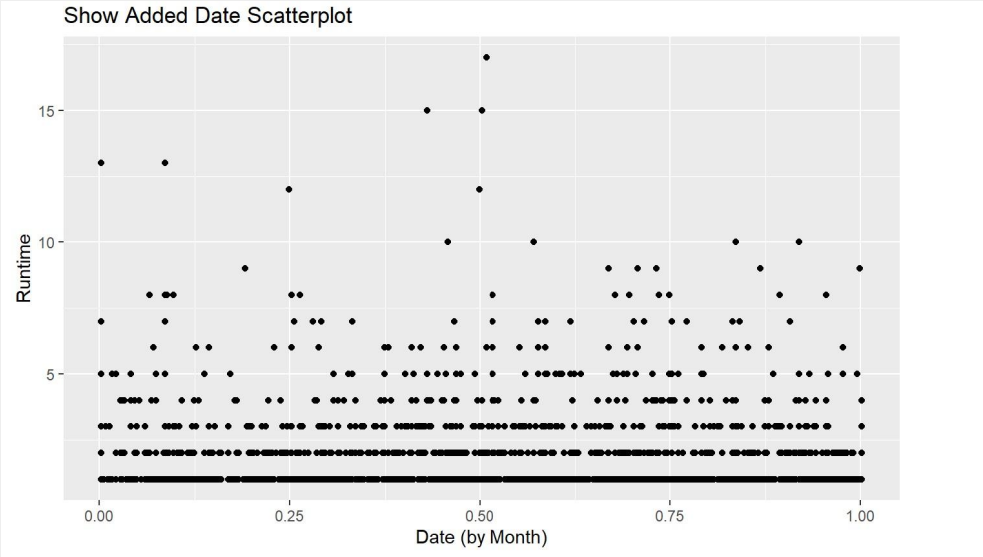


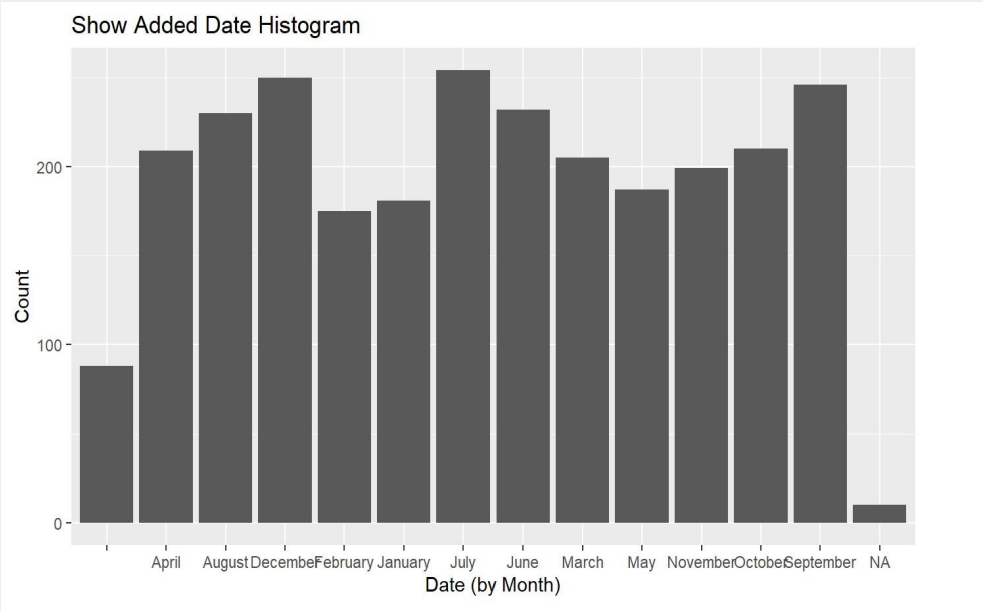


Maybe certain times of the year are meant for longer movies. Exploring this, there does not seem to be any correlation. Still there are interesting lines that appear in the scatterplot that may be indicative of certain days which more movies are added. Similarly in the Histogram, there appears to be months (July) with many more new listings than others (February).

III. C. 2. Time of Year Added to Netflix: Shows

Having looked at movies, I of course had to explore for some correlation with shows as well. Like before, there does not seem to be any correlation between these variables. But also like with movies, there does appear to be months with many more listings (July) than others (February).





IV. Summary of Initial Exploratory Analysis

Having now looked at how each of the variables corresponds (or doesn't correspond) to runtime in minutes for movies and seasons for tv shows, there appeared to be some factors that appears to strongly determine the runtimes, many factors that had no connection, and a few that could use a bit more exploration.

The factors that had the strongest correlations included the directors for movies, with the directors with the longest movies producing movies over twice as long as those with the shortest (144 min vs 64 min), the country of production for movies, with a 44 minute difference between the longest and shortest movie average per country, and genre for movies and tv shows, with a 46 minute difference for movies and a 2.1 season difference for tv shows.

The factors that had the least correlation included time of year added for movies and tv shows, date added for tv shows, and tv rating for tv shows.

Factors such as release year for movies and tv shows, date added to netflix for movies, and ratings for movies could use a bit of further exploration.

IV. Possible Considerations for Part II

As I consider continuing this project onto part II, I would like to do a lot of hypothesis testing within the factors I identified as appearing to have strong correlations to determine whether or not the observed correlations are actually statistically significant. If they are, I would like to explore those factors in reference to other factors in the dataset beyond runtime to see if I can find other interesting correlation that may serve as evidence for some possible explanations of this. I would also like to set up confidence intervals to see if the factors correlations are strong enough to make predictions of runtime given the factor.

For the factors I identified as having a weak/no correlation, I will likely not work with very much unless they come up for some reason with regards to the others.

For the factors I identified as needing more exploration, I will explore in reference to other factors beyond runtime and use methods such as linear regression on the quantitative variables to hopefully get a better sense of the data and correlations.

Thanks