Data Science exam: Question 4 (Netflix)

Erik Valentin Schulte[a]

[a]*Stellenbosch University, South Africa*

Abstract

x.

1. Introduction

Both datasets have a common identifier by which it can be knitted. I merge the two dataset and perform the following analysis. First, I present a table with some summary statistic about the newly merged dataset.

```
##
## =============================================================
## Statistic         N       Mean     St. Dev.    Min     Max
## -------------------------------------------------------------
## release_year    77,213 2,014.921     8.133    1,953   2,022
## runtime         77,213   96.494     35.540      0      251
## seasons         13,976   2.074       2.269      1       42
## imdb_score      72,937   6.466       1.106    1.500   9.500
## imdb_votes      72,850 58,719.810 155,392.800   5   2,268,288
## tmdb_popularity 77,202   27.792     67.366    0.600 1,823.374
## tmdb_score      76,093   6.685       1.026    1.000  10.000
## -------------------------------------------------------------
```

Using the stargazer function we can see that the earliest movie was released in 1953 and the latest movies in 2022, while the medians of movies releases is 2018. The maximum runtime is 251 minutes and the mean run time for each movie is 96 minutes.

*Corresponding author: Erik Valentin Schulte
Email address: 26802325@sun.ac.su (Erik Valentin Schulte)

The range of the seasons of the netflix data is 1 to 42.

IMDB scores range from 1.5 to 9.5, the average rating of all the movies contained in the dataset is 6.466.

For the tmb score, there the rating scale ranges from 1 to 10 and the average rating is slighly higher with 6.685.

2. What are good movies?

First, I looks at specific directors that did well, specifically those who got ratings from the Internet Movie Database and the TMDB score of higher than 8.4. 8.4 was choosen arbitrarily to have just the right amount of observations in the graph.

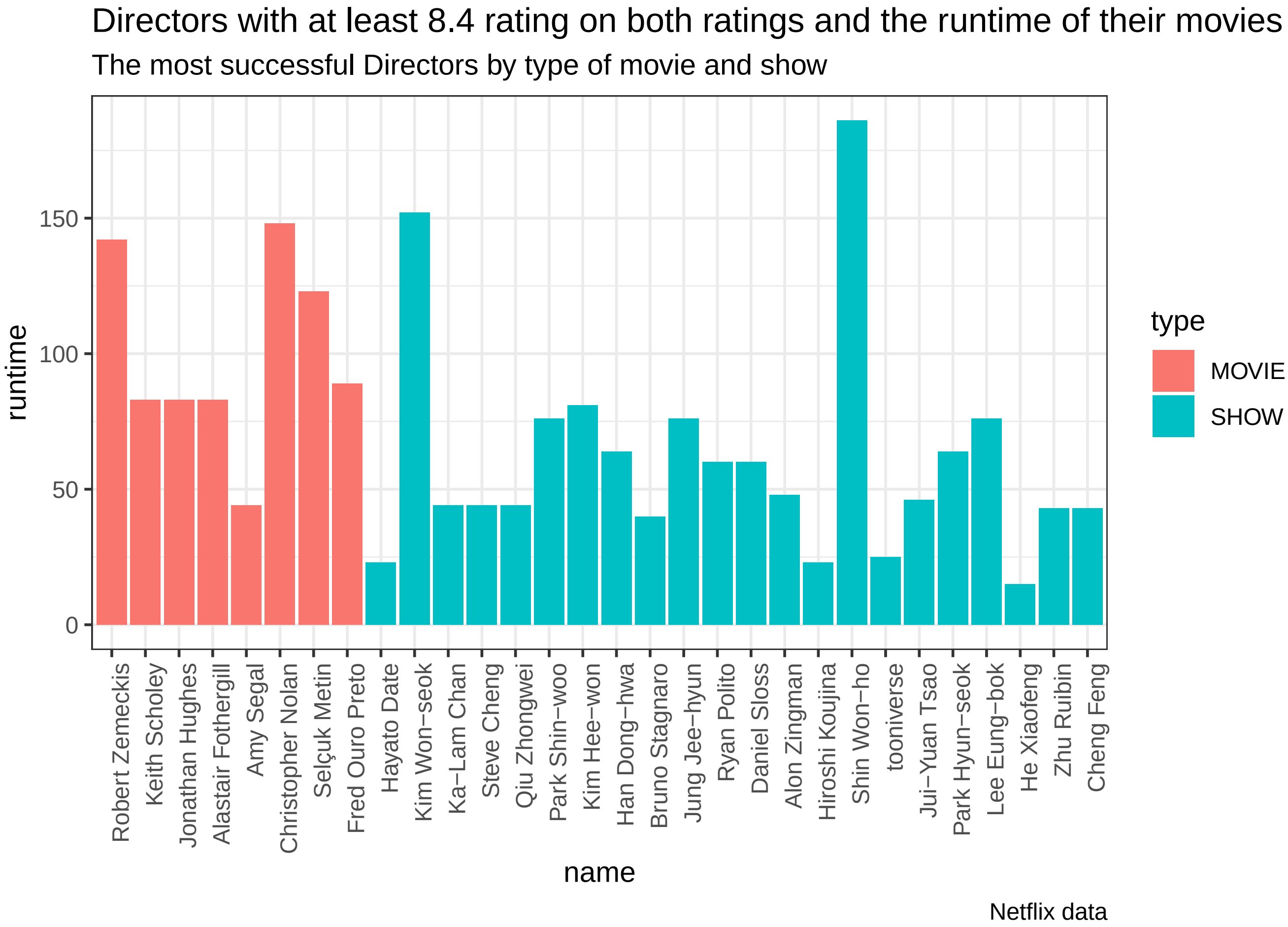

Figure 2.1: Good movies.

We can see from graph 1 that there are a few directors that did really well. There is also a lot of variation in the runtime from very short shows with less than 30 minutes by the Director He Xiaofeng or rather long shows by Shin Won-ho which take more than 150 minutes. More of the good ratings are allocated towards shows rather than movies. I would recommend my superiors into looking into the works of the Directors I found.

3. What are bad movies?

The data are also rich in what my superiors should not do. They should be careful with works that were produced by single countries only. There, seems to be especially bad movies and shows coming from India and the US, since the cumulative runtime exceeds 25000 hourse for India and 15.000hours for the US.. However, these are most likely also two very high producing countries. It is also apparent from the bad movie data that most bad movies do not have an age certificaiton. Most are available for India (PG-13) and the US (R) and Japan (TV-PG).

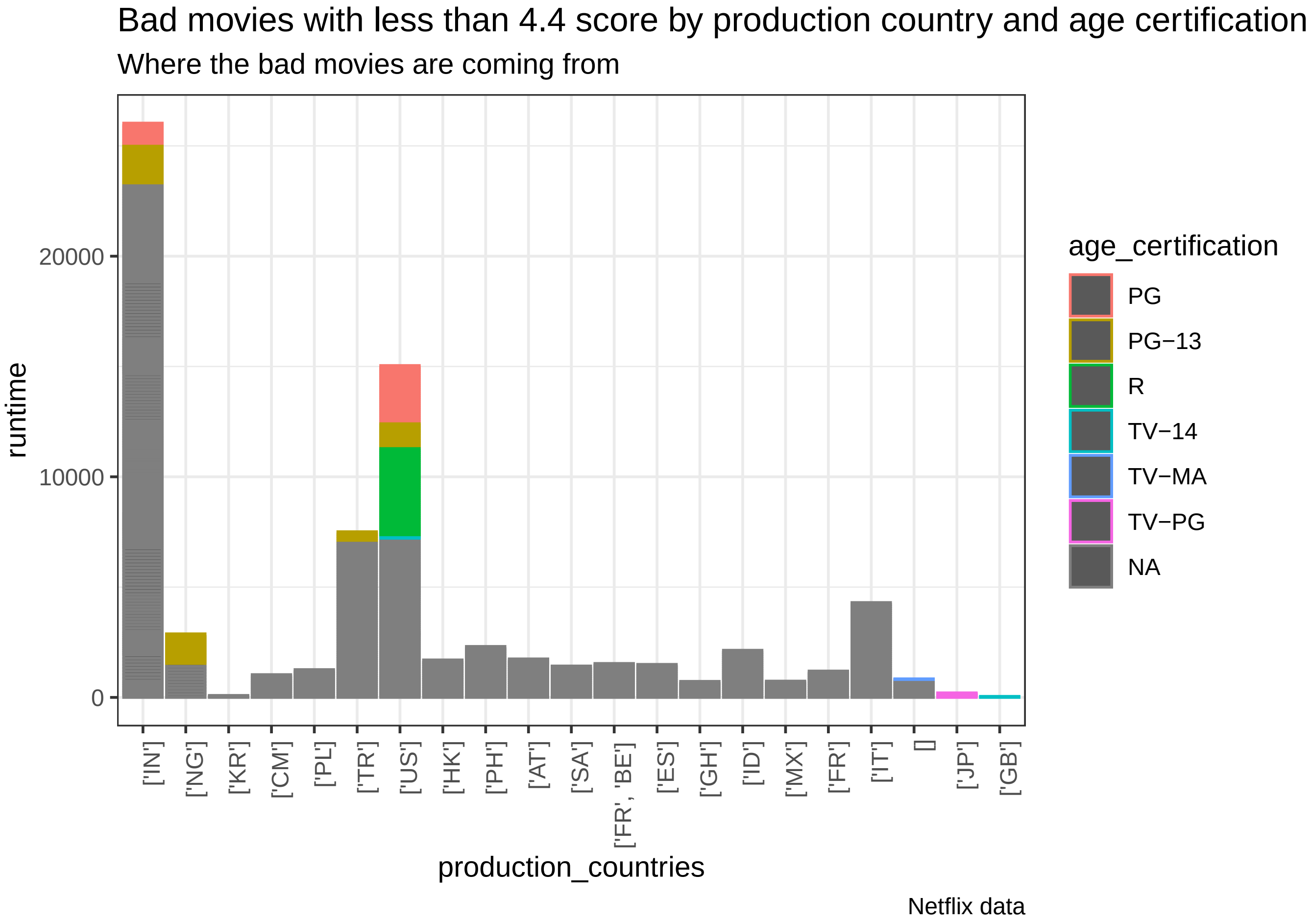

Figure 3.1: Bad movies.

Looking at the graph with the bad movies with low ratings we can see where they were produced. Interestingly, the worst movies were produced by single countries and not by co-production.