Data Overview IMDb Datasets

Group 7
Mauro de Kort, Ruben van der Thiel, Martijn Hendriks, and Sem Niezink

06-09-24

Research Motivation

The relationship between the number of episodes a TV show is set to have and its average rating is a crucial yet insufficiently studied area in the field of media research. Since competition among streaming platforms and TV networks is rising, uncovering and understanding any factor that may influence TV show rating is paramount for optimizing content. Moreover, as adult shows may benefit from having more episodes due to possibly having more complex or mature story lines, researching whether the effect of episode count on ratings differ for this genre offers additional value to this research. This study therefore aims to answer the question: "To what extent does the number of a TV show's episodes impact its average rating, and does this differ between adult titles and non-adult titles?" The insights gained from this research could assist producers in making more informed decisions with regards to episode count when creating content.

A multiple linear regression will be the applied research method, with average show rating as the dependent variable. The independent variables will consist of the continuous variable "number of episodes", as well as the dummy variable "adult title" (with 1 for adult shows, 0 for non-adult shows). By including the interaction term episodesXadult, we can also assess a potential difference in effect between adult versus non-adult movies. This linear regression method effectively addresses the objective of this research as it quantifies the impact of episode count ratings while also allowing an interaction term to assess whether this effect differs for the adult genre.

Data exploration

This report provides an overview of the 3 IMDb datasets that we are using in our research. We explore the raw data files and explain the variables to understand the structure and content of the data.

The following packages are required for this project:

```r
library(tidyr)
library(dplyr)
library(readr)
library(knitr)
library(ggplot2)
library(kableExtra)
```

Load the data files

Load the 'title basics', 'title ratings' & 'title episode' datasets.

Explanation of the data files

title.basics.tsv.gz This file contains basic information about the titles from the movies and TV shows in the IMDb database.

Table 1: Variables in title.basics

Variable	Description
tconst	Alphanumeric unique identifier of the title.
titleType	Type of title (e.g., movie, short, tvseries, tvepisode).
primaryTitle	The most popular title at the time of release.
originalTitle	Title in the original language.
isAdult	Indicates whether the title is adult content (0: No, 1: Yes).
startYear	The year the title was first released.
endYear	The year the title ended (NA for non-series).
runtimeMinutes	Runtime of the title in minutes.
genres	Includes up to three genres associated with the title.

View the first rows of the data.

```
## # A tibble: 6 x 9
##   tconst    titleType primaryTitle       originalTitle isAdult startYear endYear
##   <chr>     <chr>     <chr>              <chr>           <dbl>     <dbl>   <dbl>
## 1 tt0000001 short     Carmencita         Carmencita          0      1894      NA
## 2 tt0000002 short     Le clown et ses c~ Le clown et ~       0      1892      NA
## 3 tt0000003 short     Pauvre Pierrot     Pauvre Pierr~       0      1892      NA
## 4 tt0000004 short     Un bon bock        Un bon bock         0      1892      NA
## 5 tt0000005 short     Blacksmith Scene   Blacksmith S~       0      1893      NA
## 6 tt0000006 short     Chinese Opium Den  Chinese Opiu~       0      1894      NA
## # i 2 more variables: runtimeMinutes <dbl>, genres <chr>
```

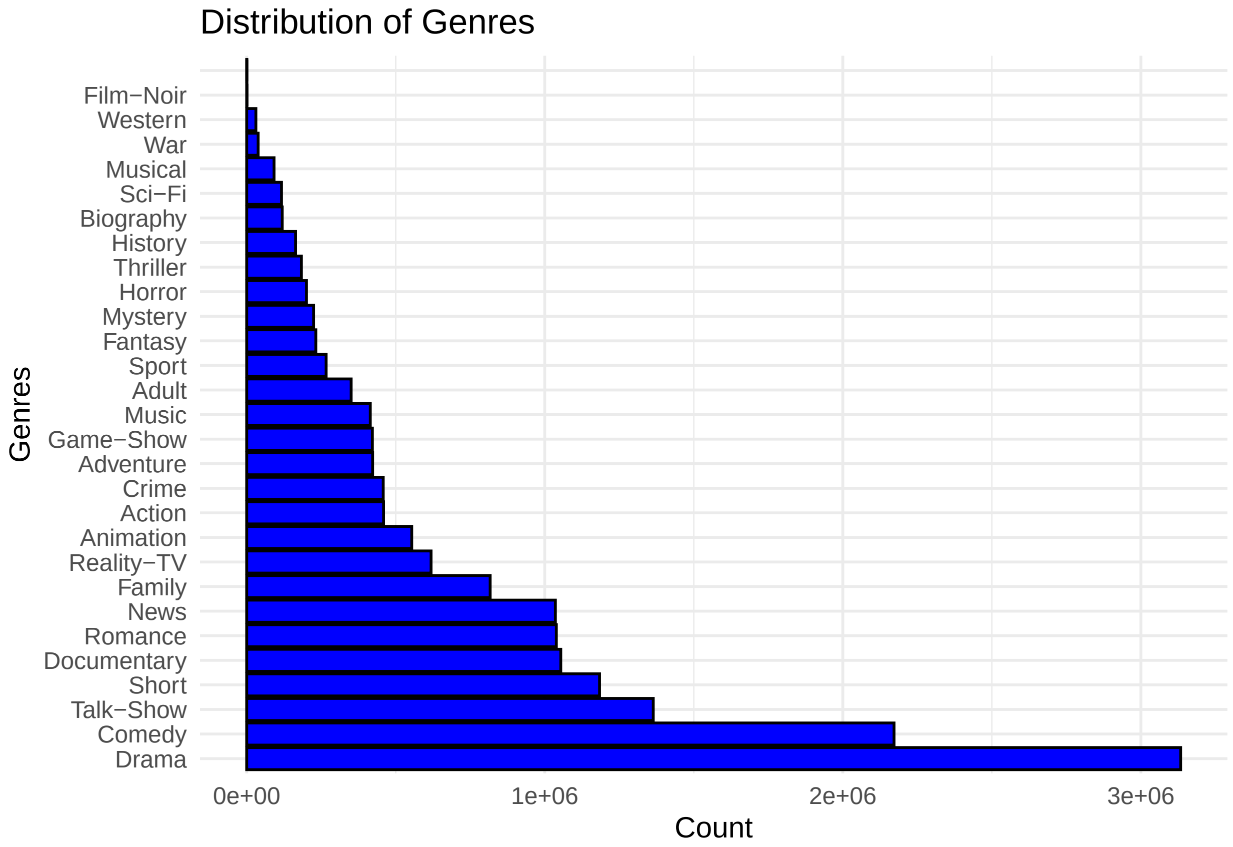

Figure 1 shows in what sizes the titles are distributed among the different genres, clearly drama and comedy are the most common.

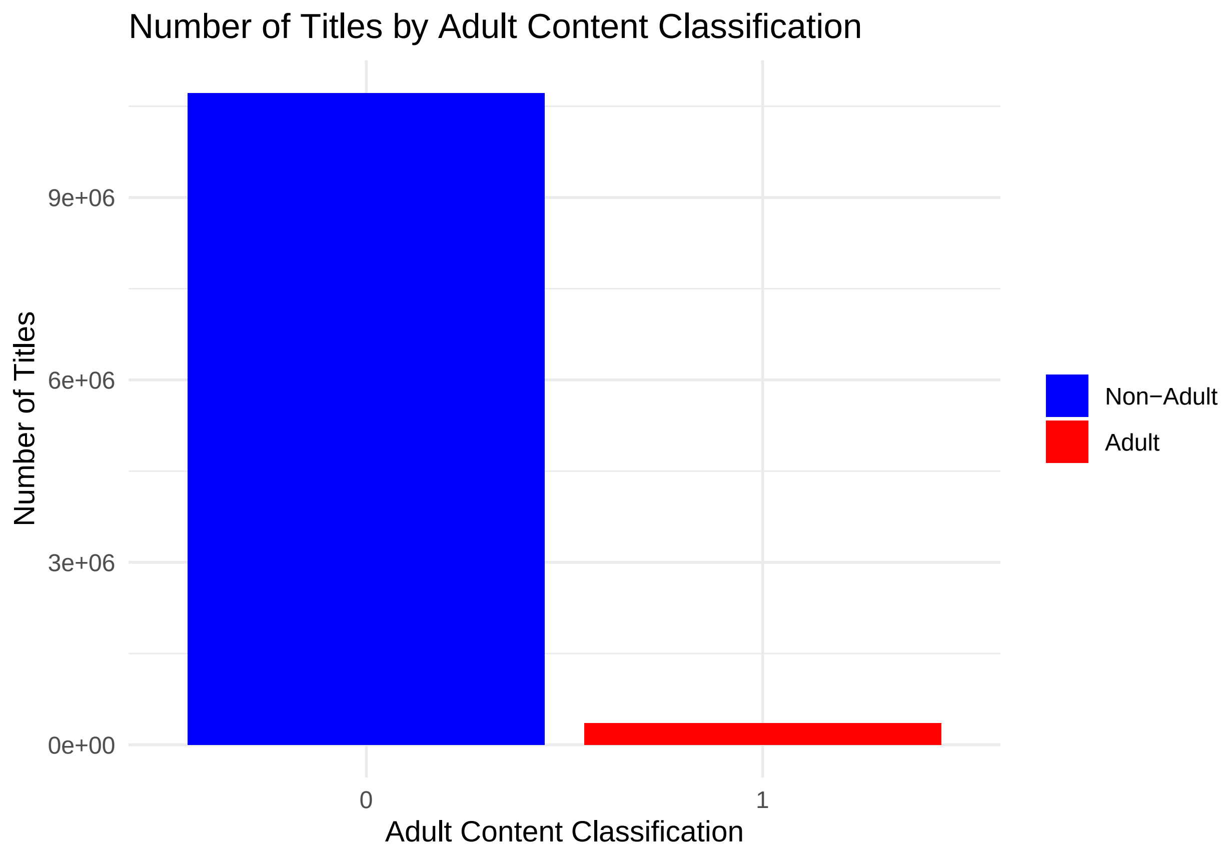

Figure 2 shows us a difference between does contain “is adult” and does not contain “is adult” titles. The vast majority of titles are not only for adults. More than 10 million titles dont have the ‘is adult’ stamp, on the other hand there are around 350,000 titles that do contain the ‘is adult’ stamp.

title.ratings.tsv.gz This file contains user ratings and the number of votes for each title.

Table 2: Variables in title.ratings

Variable	Description
tconst	Alphanumeric unique identifier of the title.
averageRating	Weighted average of all user ratings.
numVotes	Number of votes the title has received.

View the first rows of the data

```
## # A tibble: 6 x 3
##   tconst    averageRating numVotes
##   <chr>             <dbl>    <dbl>
## 1 tt0000001           5.7     2086
## 2 tt0000002           5.6      283
## 3 tt0000003           6.5     2090
## 4 tt0000004           5.4      184
## 5 tt0000005           6.2     2824
## 6 tt0000006           5        195
```

Analyse the data

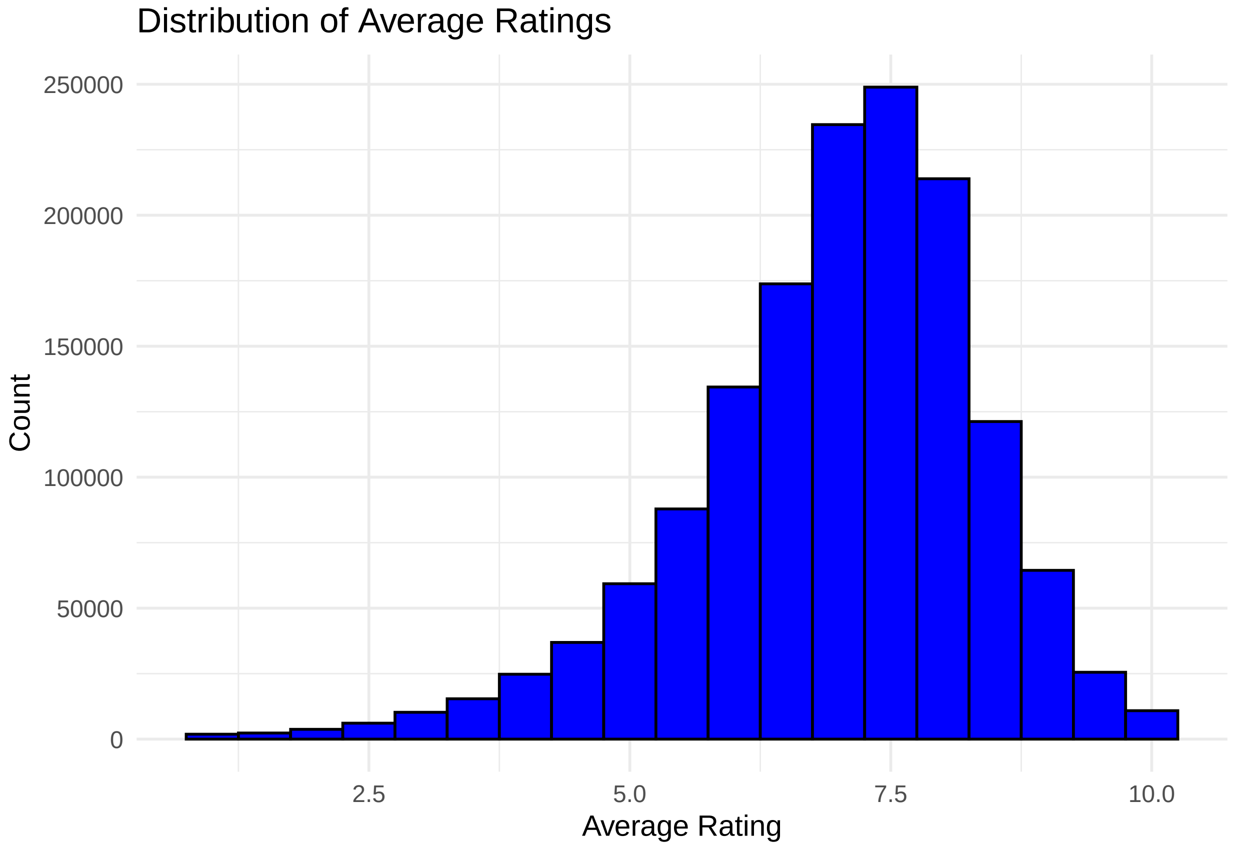

Figure 3 shows the distribution of ratings for the te titels, the highest frequency takes place between grades 6.0 and 8.0 with a peak around 7.5. Furthermore, there are fewer lower ratings for the titels.

Table 3: Table 1: Number of Titles per Voting Category

vote_category	count
0-100	1106709
101-1,000	278567
1,001-10,000	75518
10,001-50,000	10484
50,001-100,000	2149
100,001+	2791

Table 1 shows how many votes the titles received. The majority has less then 100 votes, there are about 5000 titels with more then 50.000 votes.

title.episode.tsv.gz This file contains information about TV show episodes.

Table 4: Variables in title.episode

Variable	Description
tconst	Alphanumeric identifier of the episode.

Variable	Description
parentTconst	Identifier of the parent TV series.
seasonNumber	The season number the episode belongs to.
episodeNumber	The episode number within the season.

View the first rows of the data

```
## # A tibble: 6 x 4
##   tconst    parentTconst seasonNumber episodeNumber
##   <chr>     <chr>               <dbl>         <dbl>
## 1 tt0031458 tt32857063             NA            NA
## 2 tt0041951 tt0041038               1             9
## 3 tt0042816 tt0989125               1            17
## 4 tt0042889 tt0989125              NA            NA
## 5 tt0043426 tt0040051               3            42
## 6 tt0043631 tt0989125               2            16
```

Table 5: Table 2: Number of TV Series per Episode Category

episode_category	count
1-5	71510
6-10	48305
11-20	35783
21-50	27482
51-100	12009
100+	14431

Table 6: Table 3: Summary Statistics for Number of Episodes per TV Series

Minimum	Maximum	Mean	Median
1	18593	40.55	8

Tables 2 and 3 give us clarity on how many episodes the TV series have. The dataset contains a maximum number of episodes of 18593 and the average number of episodes per TV series is 8.

References

IMDb Datasets: https://developer.imdb.com/non-commercial-datasets/