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.

First download required packages

```
library(dplyr)
library(readr)
```

Data Files

title.basics.tsv.gz

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

1. Load the 'title.basics' dataset

```
title_basics <- read_delim('https://datasets.imdbws.com/title.basics.tsv.gz', delim = '\t', na = '\\N')
```

2. View the first rows and the structure of the data

```
head(title_basics)
```

```
## # 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>
```

```
str(title_basics)
```

```
## spc_tbl_ [11,049,789 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ tconst        : chr [1:11049789] "tt0000001" "tt0000002" "tt0000003" "tt0000004" ...
##  $ titleType     : chr [1:11049789] "short" "short" "short" "short" ...
##  $ primaryTitle  : chr [1:11049789] "Carmencita" "Le clown et ses chiens" "Pauvre Pierrot" "Un bon bo
##  $ originalTitle : chr [1:11049789] "Carmencita" "Le clown et ses chiens" "Pauvre Pierrot" "Un bon bo
##  $ isAdult       : num [1:11049789] 0 0 0 0 0 0 0 0 0 0 ...
##  $ startYear     : num [1:11049789] 1894 1892 1892 1892 1893 ...
##  $ endYear       : num [1:11049789] NA NA NA NA NA NA NA NA NA NA ...
##  $ runtimeMinutes: num [1:11049789] 1 5 5 12 1 1 1 1 45 1 ...
##  $ genres        : chr [1:11049789] "Documentary,Short" "Animation,Short" "Animation,Comedy,Romance"
##  - attr(*, "spec")=
##   .. cols(
##   ..   tconst = col_character(),
##   ..   titleType = col_character(),
##   ..   primaryTitle = col_character(),
##   ..   originalTitle = col_character(),
##   ..   isAdult = col_double(),
##   ..   startYear = col_double(),
##   ..   endYear = col_double(),
##   ..   runtimeMinutes = col_double(),
##   ..   genres = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Variables

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.

title.ratings.tsv.gz

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

1.Load the 'title.ratings' dataset

```
title_ratings <- read_delim('https://datasets.imdbws.com/title.ratings.tsv.gz', delim = '\t', na = '\\N
```

2. View the first rows and the structure of the data

```
head(title_ratings)
```

```
## # A tibble: 6 x 3
##   tconst    averageRating numVotes
##   <chr>             <dbl>    <dbl>
## 1 tt0000001           5.7     2081
## 2 tt0000002           5.6      280
## 3 tt0000003           6.5     2078
## 4 tt0000004           5.4      181
## 5 tt0000005           6.2     2816
## 6 tt0000006           5        194
```

```
str(title_ratings)
```

```
## spc_tbl_ [1,472,885 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ tconst       : chr [1:1472885] "tt0000001" "tt0000002" "tt0000003" "tt0000004" ...
##  $ averageRating: num [1:1472885] 5.7 5.6 6.5 5.4 6.2 5 5.4 5.4 5.4 6.8 ...
##  $ numVotes     : num [1:1472885] 2081 280 2078 181 2816 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   tconst = col_character(),
##   ..   averageRating = col_double(),
##   ..   numVotes = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Variables tconst: Alphanumeric unique identifier of the title

averageRating: Weighted average of all user ratings

numVotes: Number of votes the title has received

title.episode.tsv.gz

This file contains information about TV show episodes.

1.Load the 'title.episode; dataset

```
title_episodes <- read_delim('https://datasets.imdbws.com/title.episode.tsv.gz', delim = '\t', na = '\\
```

2. View the first rows and the structure of the data

```
head(title_episodes)
```

```
## # 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
```

```
str(title_episodes)
```

```
## spc_tbl_ [8,472,376 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ tconst       : chr [1:8472376] "tt0031458" "tt0041951" "tt0042816" "tt0042889" ...
##  $ parentTconst : chr [1:8472376] "tt32857063" "tt0041038" "tt0989125" "tt0989125" ...
##  $ seasonNumber : num [1:8472376] NA 1 1 NA 3 2 2 3 1 2 ...
##  $ episodeNumber: num [1:8472376] NA 9 17 NA 42 16 8 3 6 16 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   tconst = col_character(),
##   ..   parentTconst = col_character(),
##   ..   seasonNumber = col_double(),
##   ..   episodeNumber = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Variables tconst: Alphanumeric identifier of the episode

parentTconst: Identifier of the parent TV series

seasonNumber: The season number the episode belongs to

episodeNumber: The episode number within the season

References

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