ProjectVersion1

Group 9

2024-09-11

Research question

Does the number of episodes significantly influence the ratings of TV shows?

Research Motivation

This research question is relevant because it allows filmmakers and marketers to understand the key factors influencing TV shows ratings, leading to a better benchmarking and decision-making in their marketing strategies. With this information, filmmakers can make better decisions in the number of episodes they are producing. Furthermore, IMDb can improve its recommendation system, offering more personalized movie recommendations, making it easier for users to find movies they will enjoy.

#The appropriate method for this research is a multiple regression analysis. The multiple regression analysis shows how several independent variables influence the dependent variable. Additionally the multiple regression is convenient to interpret. With this type of analysis it can be determined how much each independent variable influences the dependent variable ratings and which independent variable influences the ratings most significantly.

2. Data preparation & analysis

2.1 Data exploration

This section explores the IMDb datasets and provide an overview of the datasets, the definitions of the variables and figures.

Table 1: First few rows of 'title.episode' dataset

tconst	parentTconst	seasonNumber	episodeNumber
tt0031458	tt32857063	NA	NA
tt0041951	tt0041038	1	9
tt0042816	tt0989125	1	17
tt0042889	tt0989125	NA	NA
tt0043426	tt0040051	3	42
tt0043631	tt0989125	2	16

Table 2: Variable Definitions for 'title.episode' Dataset

Variable	Definition
tconst	Unique identifier for the title
parentTconst	Identifier of the parent title (e.g., the series for an episode)
seasonNumber	Season number of the episode
episodeNumber	Episode number in the seasons

Table 3: First few rows of 'title.ratings' dataset

tconst	averageRating	numVotes
tt0000001	5.7	2089
tt0000002	5.6	283
tt0000003	6.5	2094
tt0000004	5.4	184
tt0000005	6.2	2827
tt0000006	5.0	196

Table 4: Variable Definitions for 'title.ratings' Dataset

Variable	Definition
tconst	Unique identifier for the title
averageRating	Average rating of the title
numVotes	Number of votes received for the title

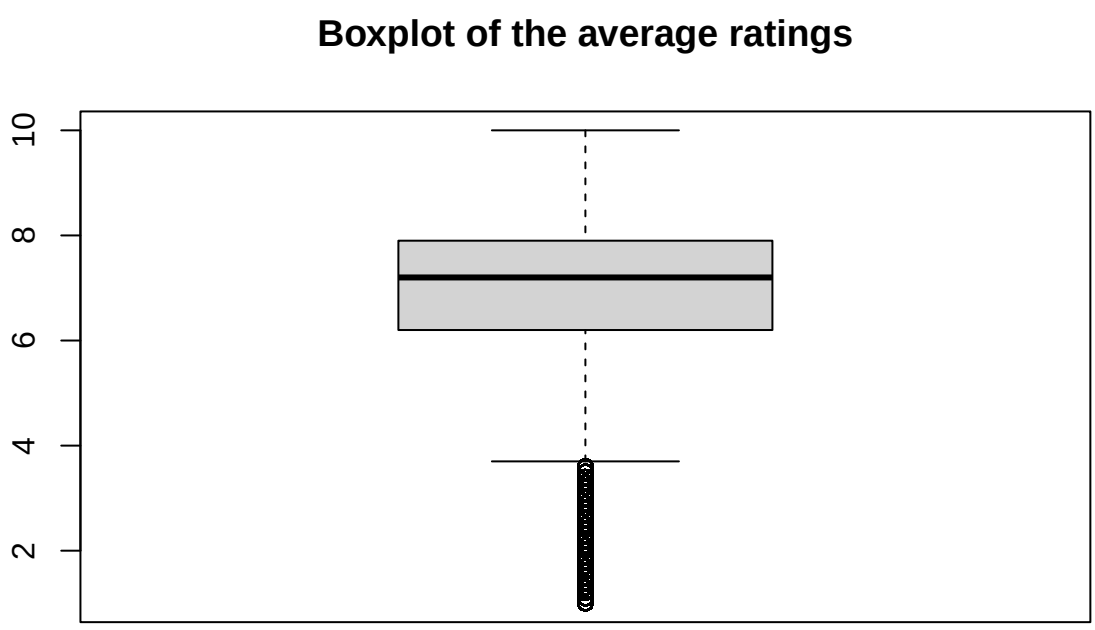

Table 5: First few rows of 'titles' dataset

tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000001	short	Carmencita	Carmencita	0	1894	NA	1	Documentary,Short
tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	NA	5	Animation,Short
tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	NA	5	Animation,Comedy,Romance
tt0000004	short	Un bon bock	Un bon bock	0	1892	NA	12	Animation,Short
tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	NA	1	Comedy,Short
tt0000006	short	Chinese Opium Den	Chinese Opium Den	0	1894	NA	1	Short

Table 6: Variable Definitions for 'titles' Dataset

Variable	Definition
tconst	Unique identifier for the title
titleType	Type of the title (e.g., movie, short, TV episode)
primaryTitle	Primary title of the work
originalTitle	Original title of the work
isAdult	Adult content flag (0: non-adult, 1: adult)
startYear	Release year or start year for series
endYear	End year for series
runtimeMinutes	Runtime in minutes

Variable	Definition
genres	Genres associated with the title

2.2 Data preparation

Since there is not a dataset available which includes the number of episodes and ratings of TV shows there needs to be new dataset created from different datasets. Before the datasets are merged the datasets needs to be cleaned. To answer the research question, the variables "Number of episodes" and "The average rating" should not contain NAs. Furthermore, some variable names are changed to make them more readable.

```
#Cleaning the episode dataset
episode = episode %>% filter(!is.na(episodeNumber))
episode = episode %>% rename("Number of the season" = "seasonNumber", "Number of episodes" = "episodeNu
```

```
#Cleaning the ratings dataset
ratings = ratings %>% filter(!is.na(averageRating)) %>% rename("The average rating" = "averageRating",
```

```
#Cleaning the titles dataset
titles = titles %>% filter(titleType == "tvEpisode")
#Renaming the headers?!?!?!
```

```
#Merging the datasets
test = right_join(episode, ratings, by = "tconst")
IMDb_dt = inner_join(test, titles, by = "tconst")
```