Data Analysis Report

Team 9

2024-10-11

1. Summary of Descriptive Statistics on Average Ratings

In this section, we present a summary of the descriptive statistics for the average ratings of TV shows. This analysis provides an overview of key metrics such as the mean, median, standard deviation, and range of ratings.

Table 1: Summary of Descriptive Statistics on Average Ratings

mean_average_rating	median_rating	min_rating	max_rating	num_episodes
7.4	7.5	1	10	729601

2. Regressing Rating on Number of Episodes

For this analysis, we decided to use regression analysis to explore the relationship between the number of episodes and the ratings of TV shows. By applying this method, we aim to determine whether the number of episodes significantly impacts a show's rating and to quantify the strength of this relationship. This approach will help us gain insights into how episode count influences audience engagement and reception.

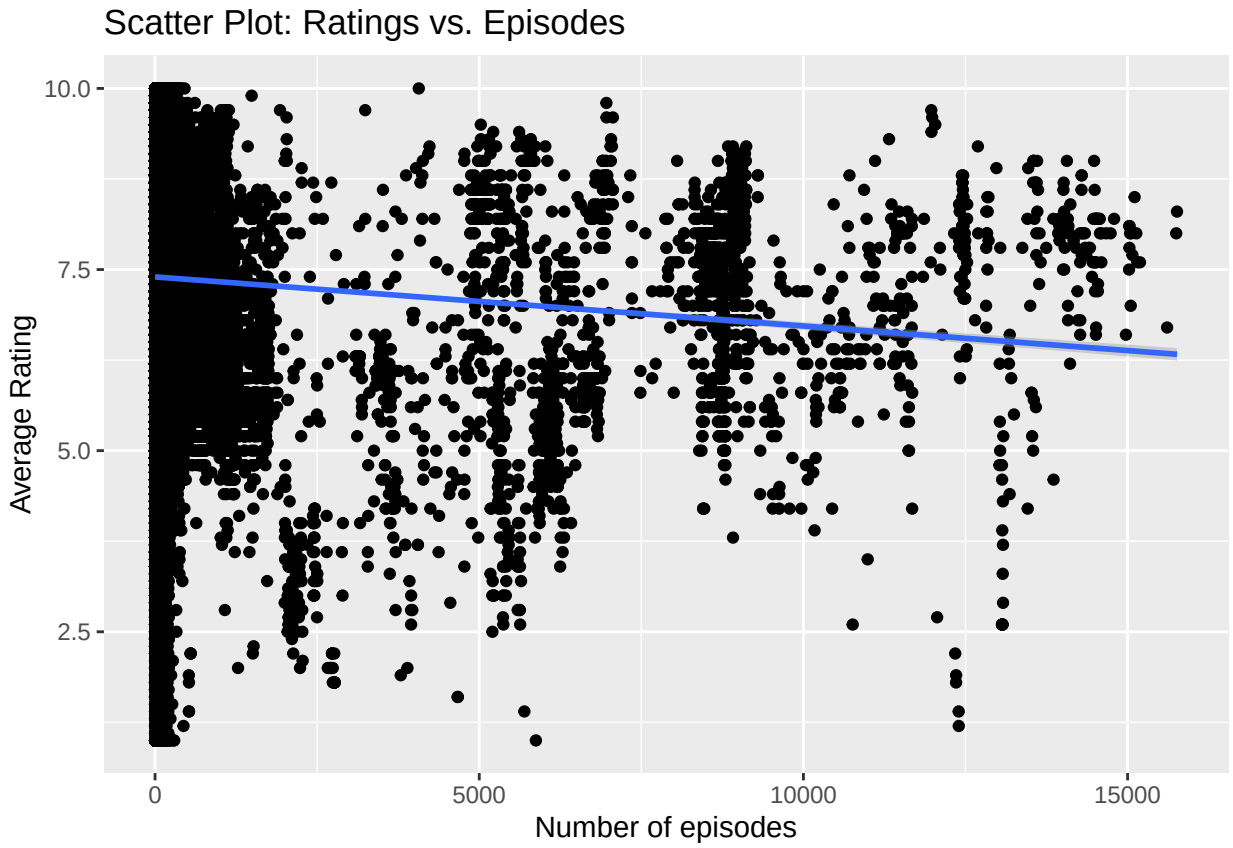
2.1 Main Model Without Control Variables

Table 2: Regression Result: Model without Control Variables

term	estimate	std.error	statistic	p.value
(Intercept)	7.3985	0.0013	5520.6726	0
Number_of_episodes	-0.0001	0.0000	-24.7156	0

Table 3: Additional Model Statistics

Statistic	Value
Coefficient for Number of episodes	-0.0001
T-value for Number of episodes	-24.7156
P-value for Number of episodes	0.0000
R-squared	0.0008
Adjusted R-squared	0.0008
F-statistic	610.8623
P-value for F-statistic	0.0000



2.2 Regression Output Analysis

In our basic model without any control variables, we observe that the number of episodes has a slightly negative effect on the average IMDb rating. With a p-value smaller than the 5% significance level, we can conclude that the number of episodes has a statistically significant negative effect. However, this model does not include any control variables, so further analysis is needed to expand and refine our model.

2.3 Main Model With Control Variables

Table 4: Regression Result: Model with Control Variables

term	estimate	std.error	statistic	p.value
(Intercept)	7.4112	0.0039	1884.6137	0.0000
Number_of_episodes	0.0000	0.0000	-2.2723	0.0231
popularity	0.6388	0.0062	102.7964	0.0000
runtimeshort	-0.0111	0.0038	-2.9466	0.0032
new_vs_oldold	-0.0273	0.0031	-8.6902	0.0000
episode_quantityMany	-0.3178	0.0048	-66.5398	0.0000

Table 5: Additional Model Statistics

Statistic	Value
Coefficient for Number of episodes	0.0000
T-value for Number of episodes	-2.2723
P-value for Number of episodes	0.0231
R-squared	0.0357
Adjusted R-squared	0.0357
F-statistic	3281.1880
P-value for F-statistic	0.0000

- popularity: "Amount of votes are over 1000
- runtime: "Runtime in minutes is more than 50"
- new_vs_old: "The start year is later than 2015"
- episode_quantity: "Number of episodes is more than 25"

2.4 Regression Output Analysis

In our main model with control variables, we observe that the coefficient for the number of episodes is negative; however, it is not significant, as the p-value for this variable is 0.772, which is greater than 0.05. Looking at our control variables, we find that all of them are significant: Popularity (amount of votes over 1000) has a significant positive effect on the average rating. In contrast, runtime (runtime in minutes is more than 50) has a significant negative effect on the average rating. Additionally, being new (the start year is later than 2015) has a significant negative effect on the average rating, and having many episodes (more than 25 episodes) also negatively affects the average rating.

3. Correlation Matrix of the Predictive Variables in Our Main Model

	Number_of_episodes	popularity	runtimeshort	new_vs_oldold	episode_quantityMany
Number_of_episodes	1.0000	-0.0173	0.0223	-0.0403	0.2172
popularity	-0.0173	1.0000	-0.0014	-0.0386	-0.0609
runtimeshort	0.0223	-0.0014	1.0000	0.0486	-0.0102
new_vs_oldold	-0.0403	-0.0386	0.0486	1.0000	0.0207
episode_quantityMany	0.2172	-0.0609	-0.0102	0.0207	1.0000

Correlation Matrix Analysis:

- Number of episodes has a moderate positive correlation (0.217) with episode quantityMany, indicating that shows with many episodes tend to be classified as having "many" episodes.
- Number of episodes has weak correlations with other variables like popularity (-0.017), runtime (0.022), and new vs old (-0.040), suggesting that the number of episodes is not strongly related to these variables.
- Popularity is weakly and negatively correlated with both new vs old (-0.039) and episode quantityMany (-0.061), indicating that neither older shows nor those with many episodes are strongly related to popularity.
- Runtime has a weak positive correlation (0.049) with new vs old, meaning that older shows may have slightly longer runtimes.

Overall, the relationships between most variables are weak, indicating little to no strong linear correlation between them.The only meaningful correlation is between Number of episodes and episode quantityMany (0.217), which makes sense as it reflects the classification of episode quantity.

4. Multicollinearity

	VIF
Number_of_episodes	1.052561
popularity	1.005158
runtime	1.003248
new_vs_old	1.006478
episode_quantity	1.054251

Multicollinearity Analysis:

All VIF values are close to 1, indicating that there is no significant multicollinearity among the variables. This suggests that each variable provides unique information to the model, and none of them are overly redundant.