ANALYSIS OF NETFLIX TV SERIES AND FILMS RATINGS BY GENRE

Jen Lung Hsu
Institute of Data Science
National Cheng Kung University
RE6121011@gs.ncku.edu.tw

December 26, 2023

ABSTRACT

This study analyzes TV series and films ratings across various genres on Netflix. Through thorough data collection, pre-processing, Anaylisis of Variance, and Post-hoc analysis, it investigates audience rating differences. The aim is to pinpoint genres with higher audience ratings, offering insights for content creation and understanding audience preferences in streaming entertainment. This research emphasizes the significance of genre-based rating disparities in shaping audience perception and engagement.

Keywords Anaylisis of Variance, Post-hoc analysis

1 Introduction

Netflix, a leading streaming platform, curates an extensive collection of TV series and films across various genres. Understanding audience preferences across these genres is pivotal for content creators to tailor offerings that captivate viewers. This study undertakes a comprehensive analysis to delve into these audience preferences.

Data processing involved meticulous selection, focusing solely on ratings with counts exceeding 100,000, ensuring a robust dataset for analysis. Recognizing that each show or movie may encompass multiple genres, the dataset underwent meticulous processing. Initially, Simple Random Sampling Without Replacement (SRSWOR) was employed to extract 20 ratings for each genre category. This step aimed to mitigate potential biases originating from ratings linked to the same TV series or films.

Subsequently, the dataset met the assumptions of normality and homogeneity of variance, prerequisites for conducting one-way Analysis of Variance (ANOVA) and post-hoc analyses. These statistical methods were employed to decipher significant differences among genres, aiming to identify the genre with the highest ratings and elucidate its prominence compared to others within the dataset. The comprehensive analyses sought to discern distinct genre preferences, offering insights into the most favored genre among audiences.

2 Statistical Method

2.1 One-Way Analysis of Variance

One-way Analysis of Variance (ANOVA) is a statistical technique used to compare means among three or more groups. It assesses whether there are statistically significant differences between the means of these groups by examining the variance within each group and between the groups. ANOVA helps determine whether the differences observed in sample means are likely to represent actual differences in the population means or if they are due to random sampling variability. This method is widely used in research to analyze the impact of categorical variables or factors on a continuous outcome or dependent variable.

In one-way ANOVA, the hypothesis is based on whether the population means of different groups are equal. The null hypotheses for ANOVA are as follows:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{1}$$

The null hypothesis states that the population means of all groups are equal, where μ_1 ,μ_2,...,μ_k represent the population means of different groups. While the alternative hypothesis claims that at least one population mean among the groups differs from the others. These hypotheses are used to assess whether there are statistical differences among the means of different groups. By testing these hypotheses, conclusions are drawn regarding whether there is sufficient evidence to suggest that at least one group's population mean differs from the others.

2.2 Post-Hoc Analysis

Post-hoc analysis conducted after ANOVA involves performing pairwise comparisons among multiple groups to determine the specific groups that exhibit statistically significant differences.

This study utilized the Tukey's method for post-hoc analysis, which compares all possible pairs of means and uses a formula to detect significant differences between group means. Here is the more detailed Tukey's HSD (Honestly Significant Difference) formula:

$$Tukey's\, HSD = q_\alpha(k, N-k) \times \sqrt{\frac{MSE}{n}} \tag{2}$$

Where:

- $q_\alpha(k, N-k)$ represents the critical value obtained from the studentized range distribution.
- α denotes the significance level.
- k is the number of groups.
- N is the total sample size.
- MSE stands for the mean square error acquired from the ANOVA table.
- n indicates the sample size.

We can utilize this formula to understand the relationship between the ratings of two different genres of TV series or films and the computed value from the Tukey's HSD formula. If the difference between the ratings of two genres exceeds the calculated value from this formula, it signifies a significant difference between these genres. Conversely, if the observed difference is smaller than this value, it indicates that there is no significant difference between them.

3 Data Collection and processing

3.1 Data Source

This dataset, titled "Netflix Engagement (Jan-Jun 23)+," was obtained from Kaggle. Here is the link:

`https://www.kaggle.com/datasets/vassyesboy/netflix-engagement-jan-jun-23`

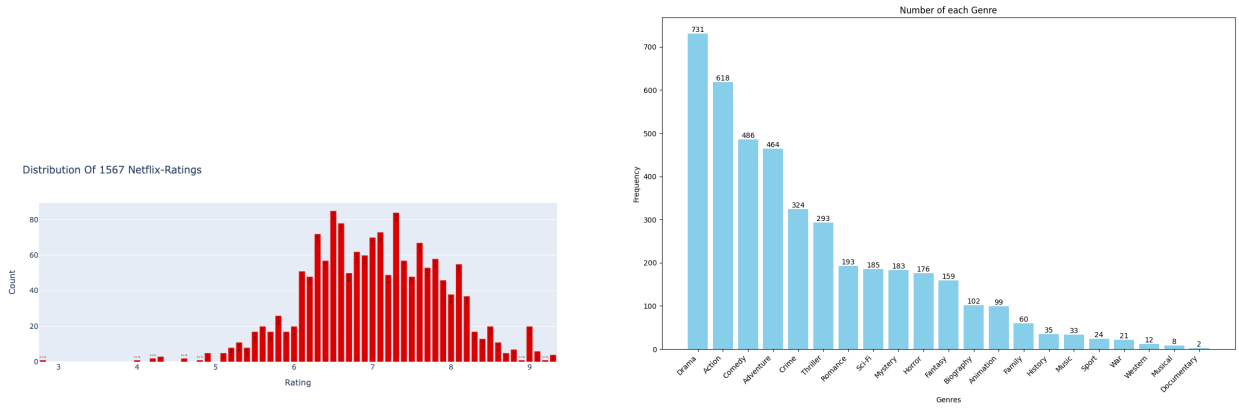
3.2 Data Preprocessing

The initial steps involved handling missing values in the 'genre' and 'ratings' fields. Rows with missing values in these columns were removed, along with the elimination of duplicate entries to ensure data integrity.

3.3 Data Processing

Under the ratings aspect, our analysis focused solely on entries exceeding 100,000 ratings. Specifically, this criterion was applied to select TV series and films for our analysis, visualizing the distribution of ratings in Figure 1a, showcasing an approximately normal distribution.

Regarding genres, given that a single show or film may encompass multiple genres, the summation of genre counts might exceed the total count of shows or films. Therefore, a bar chart was plotted to display the counts of different genres within the dataset that met the criterion of having over 100,000 ratings. Figure 1b provided an overview of the prevalence of various types of TV series or films.



(a) Histogram of Ratings

(b) Bar chart of number of each genre

Figure 1: Distribution of ratings and an overview of the prevalence of various types of TV series or films

In this section, an initial examination of the data involved extracting ratings corresponding to each genre, computing their mean, standard deviation, and quantity, followed by sorting them based on their average ratings, as shown in Table 1a. Notably, the genre 'Western' exhibited the highest average rating, whereas 'Horror' had the lowest. However, solely assessing the numerical values does not ascertain whether the ratings of one category are significantly higher or lower than others. Hence, subsequent analysis of variance was planned.

Table 1: Statistics of Ratings corresponding to each genre before and after sampling

(a) Before

Genre	Mean	std	length
Western	7.64	0.88	12
War	7.63	0.74	21
History	7.52	0.49	35
Biography	7.46	0.69	102
Animation	7.40	0.93	99
Drama	7.37	0.77	731
Musical	7.22	0.49	8
Family	7.17	0.83	60
Crime	7.16	0.82	324
Sport	7.12	0.68	24
Music	7.08	0.68	33
Mystery	7.08	0.75	183
Adventure	6.97	0.94	464
Romance	6.94	0.74	193
Thriller	6.90	0.81	293
Documentary	6.90	0.42	2
Comedy	6.89	0.78	486
Action	6.83	0.88	618
Sci-Fi	6.74	0.86	185
Fantasy	6.72	0.89	159
Horror	6.59	0.73	176

(b) After

Genre	Mean	std	length
War	7.60	0.73	20
History	7.54	0.49	20
Family	7.40	0.65	20
Drama	7.40	0.74	20
Animation	7.30	0.96	20
Music	7.25	0.66	20
Biography	7.23	0.80	20
Mystery	7.21	0.70	20
Adventure	7.12	0.88	20
Sport	7.10	0.52	20
Crime	7.04	0.88	20
Romance	7.00	0.70	20
Comedy	6.91	0.59	20
Thriller	6.85	0.76	20
Action	6.79	0.84	20
Fantasy	6.66	0.99	20
Sci-Fi	6.63	0.72	20
Horror	6.47	0.74	20

An inherent challenge arose due to TV series or films typically spanning two or three genres. To counter this, simple random sampling of 20 ratings per category was performed, serving as representative samples. This approach aimed to mitigate instances where ratings for two categories might correspond to the same TV series or films. Additionally, categories with fewer than 20 samples were excluded. This elimination criterion was based on the assumption that categories with scant representation likely did not garner over 100,000 ratings universally. The statistics of ratings of

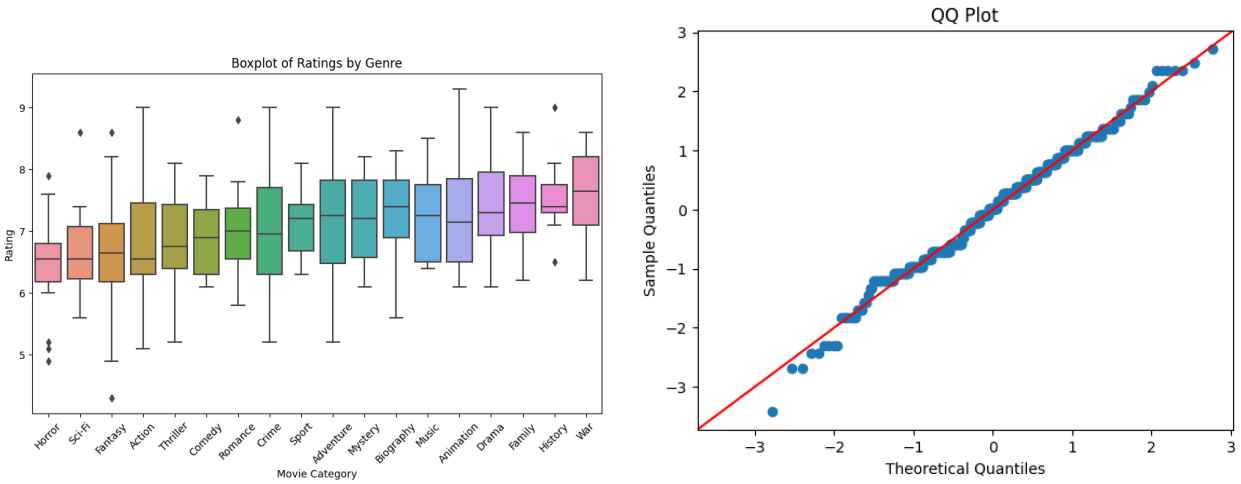
each genre is showed in Table 1b The analysis sought to encompass genres favored by audiences while considering TV series or films with the highest ratings, aiming for more factual insights.

4 Statistical Analysis

To prepare for subsequent analysis of variance, homogeneity of variance and normality tests were conducted.

4.1 Homogeneity of Variance Test

Both the Levene's test and Bartlett's test were conducted and are presented in Table 2a, with resulting p-values exceeding 0.05. This suggests homogeneity of variance among genres, allowing us to consider their variances as equal. A boxplot graph displaying the ratings for each genre was also generated in Figiure 2a.



(a) Box plot of ratings of each genre

(b) QQ plot

Figure 2: Visualization of model diagnostics

4.2 Normality Test

Utilizing the Shapiro-Wilk test and Lilliefors test, which are presented in Table 2b, p-values exceeding 0.05 in the Shapiro-Wilk test indicate that the ratings conform to a normal distribution. Additionally, a Quantile-Quantile plot (QQ plot) was created in Figure 2b, indicating proximity to the 45-degree line. These tests provide the necessary assumptions and diagnostics for conducting subsequent ANOVA analysis on the genre ratings.

Table 2: Model Diagnostics

(a) Homogeneity of Variance Test

Test	Statistic	P-value	result
Levene's test	1.0380	0.4158	do not reject
Bartlett's test	21.8148	0.1919	do not reject

(b) Nornmality Test

Test	Statistic	P-value	result
Shapiro-Wilk test	0.9922	0.0564	do not reject
Lilliefors test	0.0704	0.0010	reject

4.3 One-Way Analysis of Variance

Upon fulfilling the aforementioned assumptions, ANOVA analysis was conducted. In Table 3, the ANOVA table revealed a p-value below 0.05, indicating at least one genre significantly differs from the others. Subsequently, a post-hoc analysis was performed to identify specific genres exhibiting significant rating differences.

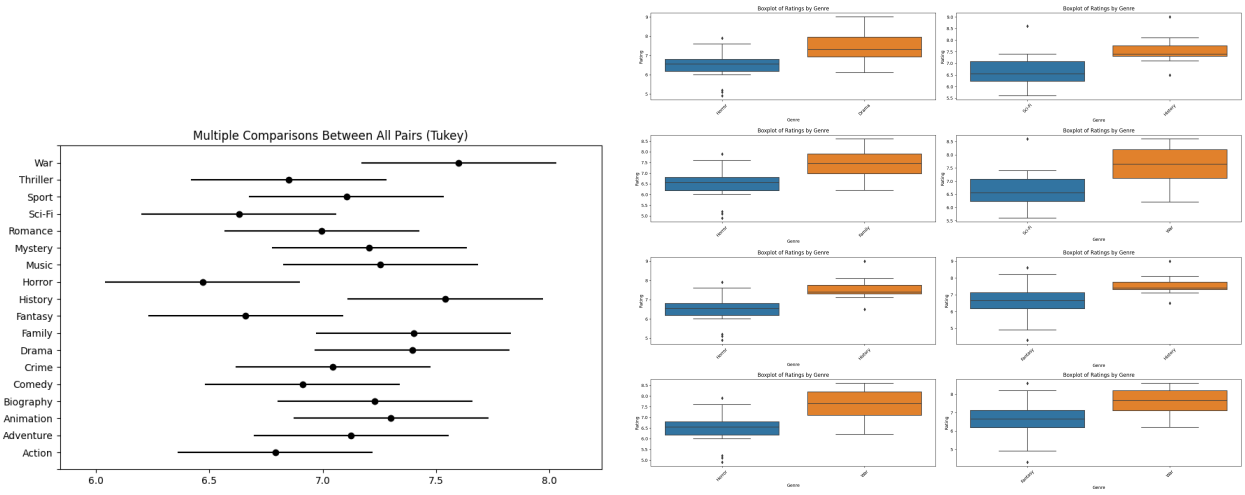
Table 3: ANOVA table

-	Sum of Square	df	F	Pr(>F)
C(genre)	34.5708	17	3.3979	0.0000
Residual	204.6825	342	-	-
Total	239.25	359	-	-

4.4 Post-Hoc Analysis

Utilizing Tukey's post-hoc test to examine particular genres with notable rating differences, the visual representation of the results is depicted in Figure 3a. In pairwise comparisons, notable differences were observed:

- 'Horror' vs. 'Drama'
- 'Horror' vs. 'Family'
- 'Horror' vs. 'History'
- 'Horror' vs. 'War'
- 'Sci-Fi' vs. 'History'
- 'Sci-Fi' vs. 'War'
- 'Fantasy' vs. 'History'
- 'Fantasy' vs. 'War'



(a) Tukey's post-hoc analysis

(b) Specific genres with significant rating discrepancies

Figure 3: Genres with Significant Rating Differences

As illustrated in Figure 3b, a closer examination of the previous graph (Figure 2a) showcased significant differences between the top four highest-rated genres and the lowest-rated genre. Additionally, the top two highest-rated genres exhibited significant differences when compared to the genres ranked second to last and third to last. Therefore, it was deduced that genres from the first to the fourth-to-last positions exhibited similarities and could be considered as a similar group. Consequently, the conclusion drawn is as follows:

Viewing in conjunction with Table 4, from 'War' to 'Action' constitutes the most favored category among viewers, representing the top-ranking genres. 'Fantasy' to 'Horror' stands as the second-ranking group, displaying ratings relatively lower compared to the top-ranked genres.

Table 4: Popular genres ranked by ratings

rank	genres
Top ranking	War, History, Family, Drama, Animation, Music, Biography, Mystery, Adventure, Sport, Crime, Romance, Comedy, Thriller, Action
Second ranking	Fantasy, Sci-Fi, Horror

5 Conclusion

Through comprehensive analytical methodologies, this study successfully discerned notable differences in ratings across various genres within the Netflix dataset. While no single genre emerged as overwhelmingly superior, the study identified clusters of genres that consistently received higher ratings compared to others.

These findings hold significant implications for the entertainment industry. By acknowledging the preferences and trends reflected in higher-rated genres, content creators and streaming platforms like Netflix can strategically tailor their offerings to cater to audience preferences more effectively. Ultimately, this study serves as a valuable foundation for further investigations and strategic decision-making in the realm of content creation and audience engagement within the streaming entertainment landscape.

6 Discussion and Future Directions

While this study provides valuable insights into Netflix ratings by genre, certain limitations persist. The absence of a definitive standout genre with significantly higher ratings might be attributed to the sampling size. A larger sample size could potentially yield more distinct differences among genres during post-hoc comparisons. However, this expansion might compromise the assumptions of homogeneity of variance and normality required for ANOVA analysis.

Moreover, the effectiveness of simple random sampling without replacement in avoiding repetitions of ratings from the same TV series or film remains a point of skepticism. This approach, while relatively straightforward and generalizable, raises concerns about potential selection biases or repeated inclusions.

Also, the p-value obtained from the normality test in this study is quite close to 0.05, indicating the possibility of considering non-parametric methods like the Kruskal-Wallis test in future analyses to address normality assumptions.

All these challenges present significant complexities in this study. Future research endeavors should consider expanding sample sizes cautiously, balancing the need for more substantial data for conclusive results while maintaining statistical assumptions. Additionally, exploring alternative sampling techniques and incorporating diverse influencing factors could enhance the robustness and applicability of findings in similar analyses.