Parametric and Nonparametric Analysis of TV Show Voting Averages

Rishi Bhakta, Andrew Guerrero, Alan Wang

Summary

This study investigated the relationship between TV show genres and their respective episode ratings, with the goal of identifying patterns that might reveal whether certain genres consistently outperform others in terms of viewer reception. Specifically, both parametric and nonparametric statistical methods were applied to compare rating distributions across genres.

While parametric methods like ANOVA and Tukey tests identified significant differences in mean ratings, nonparametric approaches such as the Kruskal-Wallis and Wilcoxon tests provided deeper insights by accounting for the non-normality and overlaps inherent in the dataset. Both parametric (ANOVA & Tukey) and nonparametric (Kruskal-Wallis, Kolmogorov-Smirnov, & Conover) tests consistently detected significant differences in TV show ratings across genres, rejecting the null hypothesis that all genres share the same rating distribution. Similarly, pairwise comparisons using Tukey and Wilcoxon Rank Sum tests identified overlapping results for many genre pairs. For instance, both methods agreed on statistically significant differences between genres such as Comedy and Drama, as well as Documentary and Action & Adventure.

These findings provide valuable insights into how TV show genres influence audience reception, with significant differences observed across genres. Additionally, from a student perspective, the consistency between parametric and nonparametric results not only validates the reliability of the analysis, but also underscores the importance of selecting appropriate methods tailored to the dataset's unique characteristics.

Table of Contents

I. Purpose of Study

The entertainment industry, particularly television, has seen a rapid expansion in content offerings across various genres, with platforms competing for audience attention. Understanding what drives audience engagement and how ratings vary across genres can provide valuable insights for content creators, producers, and distributors. This study aims to explore the relationship between TV show genres and their episode ratings, focusing on identifying whether certain genres consistently outperform others in terms of viewer reception.

The motivation for this analysis stems from the increasing reliance on data-driven decision-making in the entertainment industry. Ratings are a key metric of audience engagement and success, influencing everything from content development to marketing strategies. By leveraging a large dataset of TV shows, this study seeks to uncover genre-specific trends and patterns, offering actionable insights into audience preferences.

From an academic perspective, this study provides an opportunity to apply statistical methods to a real-world dataset, allowing for a comparison of parametric and nonparametric approaches. This study emphasizes the importance of selecting appropriate methods based on the data's characteristics, such as non-normality and overlapping groups, to derive robust and meaningful conclusions. Additionally, it highlights how nuanced analysis can reveal deeper insights that may otherwise be overlooked, contributing to the broader understanding of consumer behavior in the television industry.

II. Data Description

The dataset being used originates from the "Full TMDb TV Shows Dataset 2024" on Kaggle, comprising 150,000 television shows and over 29 variables. These variables capture key details of TV shows broadcasted between January 1, 2024, and November 8, 2024. The dataset provides a comprehensive overview of each show, including episode and season counts, genres, language, descriptions, airing dates, vote counts, and average ratings, offering a rich source of information for analysis.

Out of all columns in the dataset, only the following listed were relevant to this study and kept.

Variable	Description
id	Unique Identifier of a TV Show
name	Original name/title of the TV show.
number_of_seasons	Number of total seasons
number_of_episodes	Number of total episodes
original_language	Language TV Show was produced in
vote_count	Total number of votes for the TV show, ranking from 1-10. If 0, no voting score exists
vote_average	Average Voting Score of a TV Show
vote_popularity	Popularity Score of a TV Show, unbounded. If 0, no popularity score exists.
genres	List of comma-delimited genres associated with the TV show

The first step in processing the data was by normalizing all listed genres for a single observation. In the original dataset, these genres were listed in one single "genres" column, and were delimited by commas. By developing code to isolate all possible genre-delimited columns, the following possible genres are isolated for 168,639 Entries:

- Action & Adventure
- Animation
- Comedy
- Crime
- Documentary
- Drama
- Family
- History
- Kids
- Music
- Musical
- Mystery
- News
- Reality
- Romance
- Sci-Fi & Fantasy
- Soap
- Talk

- War & Politics
- Western

Now that the data has been normalized, the data is then cleaned based off the following rules in the following order:

1) English-Only TV Shows: The scope of this study is limited to English-language TV shows; entries in other languages are excluded.
    - *Reduces to 76,304 Entries*
2) Unique Entries: Duplicate entries are removed to ensure uniqueness. This addresses cases where the same TV show appears multiple times due to differences in TV channel providers.
    - *Reduces to 75,741 Entries*
3) Null Values: Remove any entries that have missing data for the 9 key variables isolated as relevant to this study. If data is missing from any of these columns, the entry is unusable
    - *Reduces to 36,856 Entries*
4) "Well-Established" TV Shows: Only TV shows meeting the following criteria are included:
    - At least 10 episodes.
        - *Reduces to 16,853 Entries*
    - A vote count is at least 5, indicating at least five viewer ratings.
        - Reduces to 6,604 Entries
    - A popularity score is greater than zero, indicating a popularity score exists
        - Reduces to 6,600 Entries
5) “Well-Established” Genres: Genres with over 600 Existing TV Shows
    - Genres Retained: Action & Adventure (983), Animation (1266), Comedy (2531), Crime (654), Documentary (647), Drama (2232), Family (612), Reality (744), Sci-Fi & Fantasy (976)
    - Genres Removed: History (0), Kids (555), Music (1), Musical (1), Mystery (536), News (81), Romance (1), Soap (49), Talk (170), War & Politics (72), Western (94)

III. Potential Data Cleaning Assumptions/Pitfalls

Genre Correlation and Redundancy: Some genres, such as 'Drama' and 'Sci-Fi & Fantasy,' are often correlated or overlap. In this study, genres were treated as independent groups to simplify the analysis. However, this approach does not account for multicollinearity, which could have skewed results and inflated the apparent importance of redundant features. This decision was made to ensure a straightforward methodology for initial insights.

"Well-Established" TV Show Criteria: "Well-established" shows were defined using thresholds (e.g., at least 10 episodes, a vote count of 5) to focus on TV shows with a meaningful level of audience engagement and ratings. While practical, these criteria are somewhat arbitrary and may exclude newer shows or niche productions that could provide valuable insights into emerging trends or cultural impacts.

"Well-Established" Genre Removal: To focus on genres with substantial representation, we excluded genres with fewer than 600 shows. This decision ensures the analysis highlights prevalent genres, but it introduces bias against less common genres, which may still hold unique and important insights.

Overall Dataset Reduction: The data cleaning steps reduced the dataset from 150,000 entries to 6,600, narrowing the focus to high-quality, relevant data. While this process was necessary to create a manageable and meaningful dataset, it limits the generalizability of this study's findings to the broader TV landscape.

Temporal Coverage: The dataset includes TV shows broadcast only between January 1, 2024, and November 8, 2024. This narrow time frame allows for a focused analysis of recent trends but restricts this study's ability to capture long-term patterns or historical insights.

IV. Exploratory Data Analysis

First, the goal is to determine whether parametric or nonparametric methods would be most appropriate for our data. There are two variables in our data that could be used to quantify the rating of a TV show: **vote_average** and **popularity**. The two response variables are visualized in consideration as histograms (see Fig. 1 and Fig. 2 below).

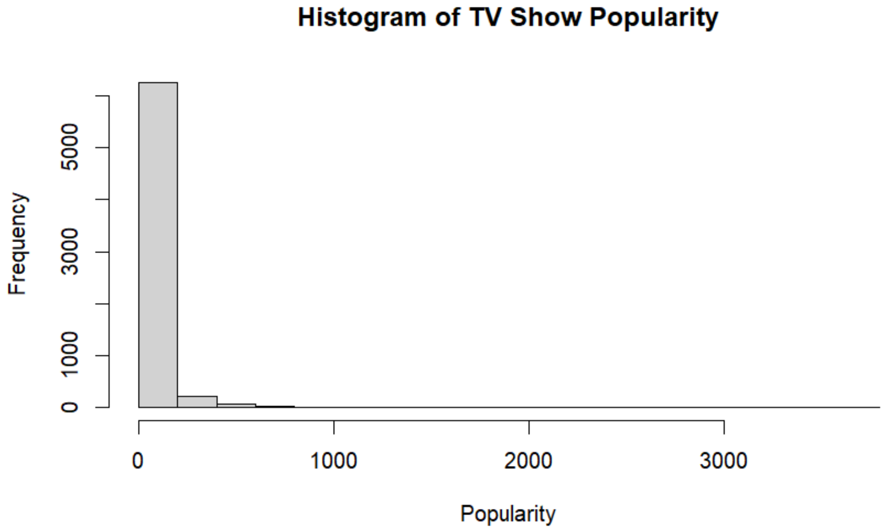

Figure 1: Histogram of Popularity Variable

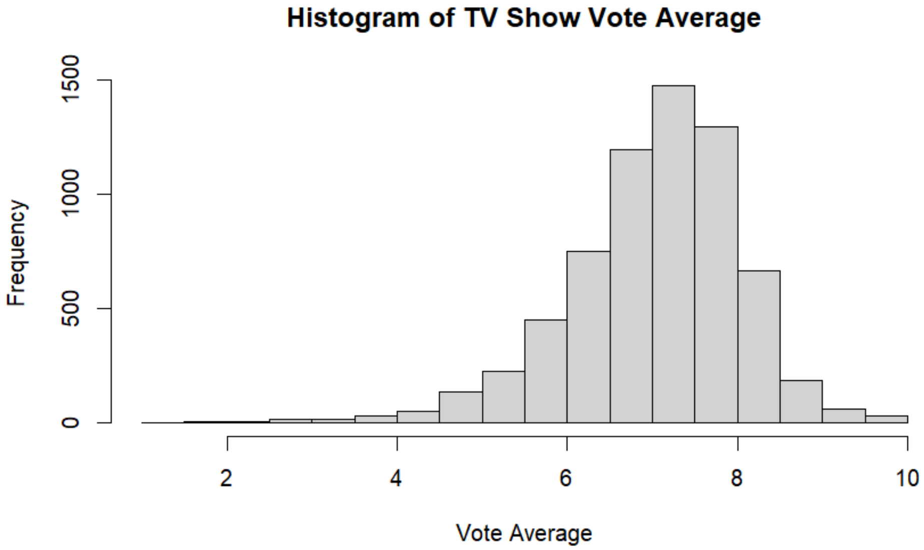

Figure 2: Histogram of Vote_Average Variable

The histograms reveal that the vote_average variable more closely follows a normal distribution compared to the popularity variable. As a result, the vote_average variable is selected as the response variable, and the popularity variable is excluded from the analysis. To test the normality of the vote_average variable, an Anderson-Darling test is conducted, given that the parameters of the distribution are unknown.

From the software analysis, the test statistic is found to be 51.38, with a p-value of approximately 0. Based on this result, the null hypothesis is rejected, indicating that the response variable is not normally distributed. Consequently, it is deemed more appropriate to utilize nonparametric methods for the analysis. However, a parametric analysis is also performed to serve as a comparison against the results obtained from nonparametric methods.

Additionally, boxplots of the vote_average variable are created for each genre to visualize the distribution of vote averages across different genres (see Fig. 3 below).

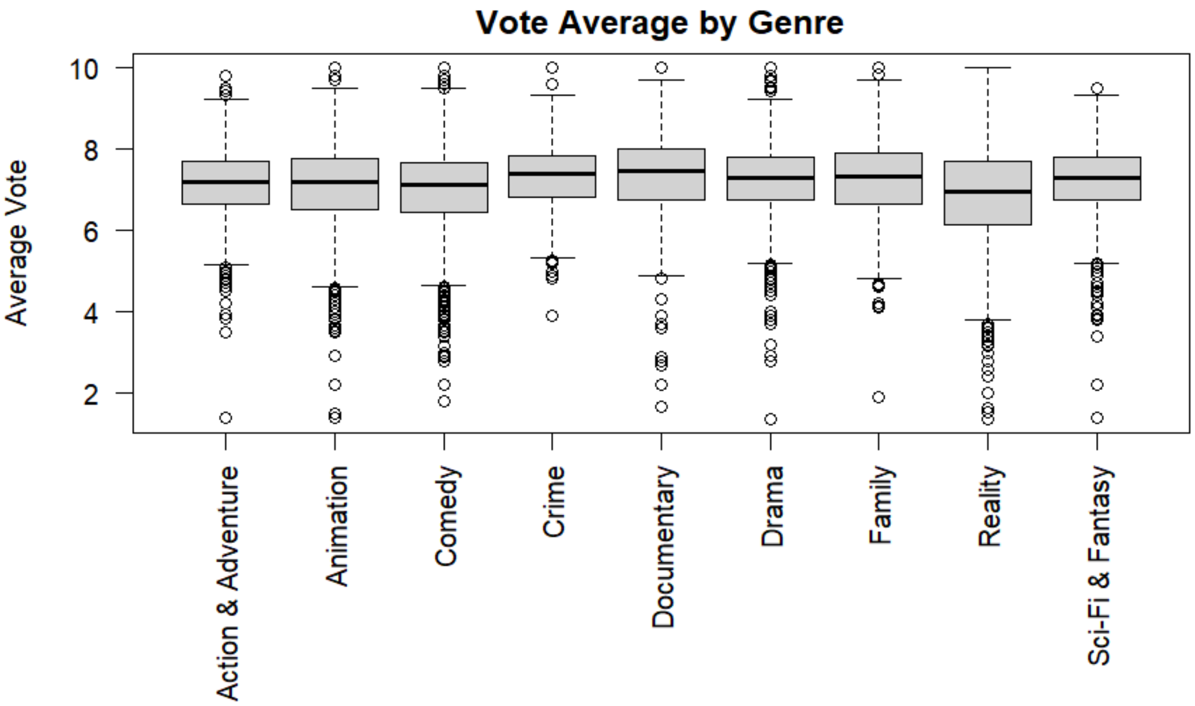

Figure 3: Histogram of Vote_Average Variable per Genre

Now, each genre's vote_average variable is adjusted for normality using Anderson-Darling tests.

H_0 : The distributions of the genre is normal.

H_1 : The distributions of the genre is not normal.

For all 9 genres, the resulting p-values are less than 0.000, so H_0 is rejected, meaning that none of the genre's response variable data are remotely normal. Thus, performing nonparametric tests on the data is much more appropriate than parametric tests.

V. Parametric Analysis

A. ANOVA

To determine whether there are differences in the means across genres, ANOVA will be used, despite the genre's data being non-normal.

H_0 : The means of the distributions of all genres are the same.

H_1 : The means of the distributions of all genres are not the same.

Using software, the test statistic is 22.27, resulting in a p-value of approximately 0, on 8 degrees of freedom (9 genres). Therefore, H_0 is rejected, i.e. there *are* differences in the means across genres. Now that it has been determined that there are differences, a post-hoc Tukey pairwise comparisons of means test is performed.

H_0 : The means of the distributions of the pair of genres are the same.

H_1 : The means of the distributions of the pair of genres are not the same.

Fig. 4 below shows the results of each pairwise comparison, with the values in each cell representing the adjusted p-value from each test. The dark red cells, with adjusted p-values less than 0.05, represent pairs of genres whose distribution means are statistically different from each other.

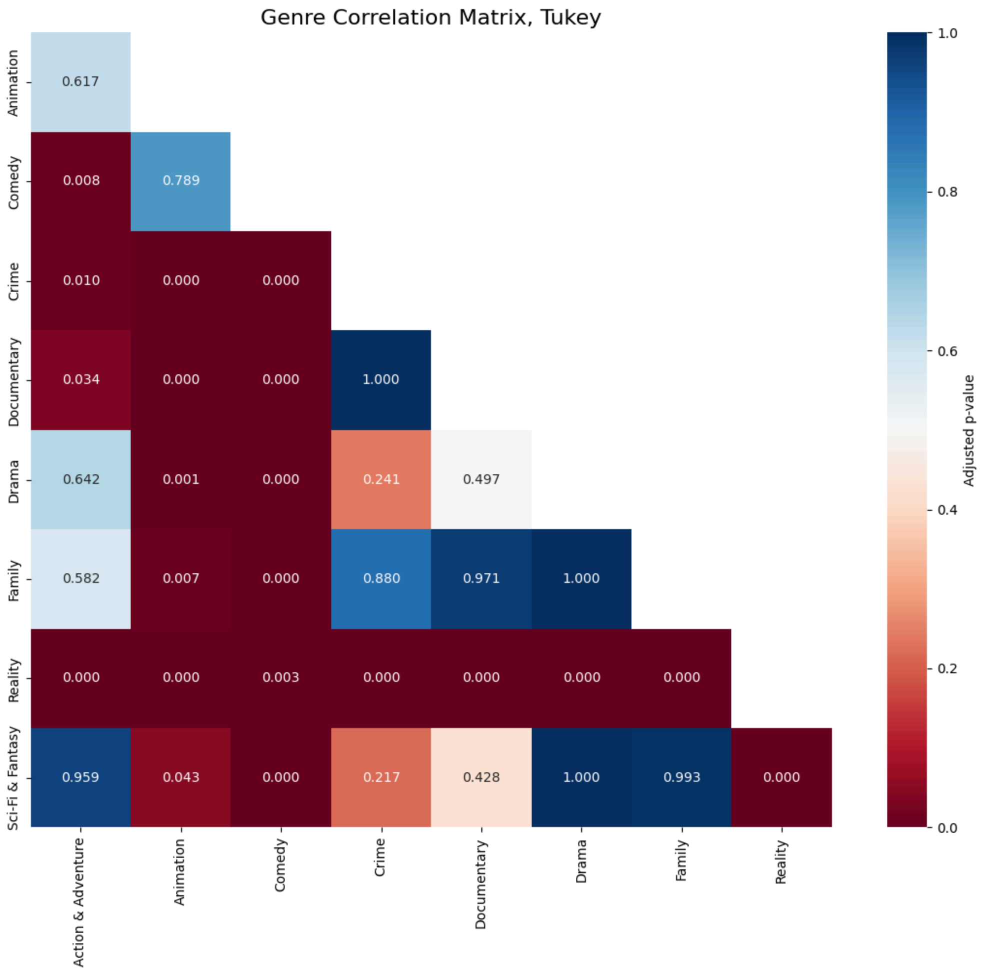

Figure 4: Genre Correlation Matrix from Tukey Test

VI. Nonparametric Analysis

A. Kruskal-Wallis Test

First, test if there are differences between a single genre and all other genres, collectively. This is done by performing a Kruskal-Wallis test for each genre using significance level α = 0.05.

H_0 : The median of the distribution of the chosen genre is the same as the median of the distribution of all other genres, collectively.

H_1 : The median of the distribution of the chosen genre is different from the median of the distribution of all other genres, collectively.

For the Animation genre, the test resulted in a p-value of 0.609, indicating that H_0 cannot be rejected, i.e. its distribution median is statistically the same as the collective distribution of all other genres. For the rest of the genres, with p-values less than the significance level of 0.05, H_0 is rejected, i.e. the median of the distribution of each genre is statistically different from the collective distribution of all other genres.

B. Wilcoxon Rank Sum Test

Given the observed differences between the medians of many genres' distributions, pairwise comparisons are conducted using a pairwise Wilcoxon Rank Sum test. This test is used to identify which pairs of genres have statistically different medians.

H_0 : The medians of the distributions of the pair of genres are the same.

H_1 : The medians of the distributions of the pair of genres are not the same.

Fig. 5 below shows the results of each pairwise comparison, with the values in each cell representing the adjusted p-value from each test. The dark red cells, with adjusted p-values less than 0.05, represent pairs of genres whose distribution medians are statistically different from each other. Once the pairs of genres that are statistically different are determined, their medians can be directly compared to conclude which genres' distribution medians are greater (or less than) the others'.

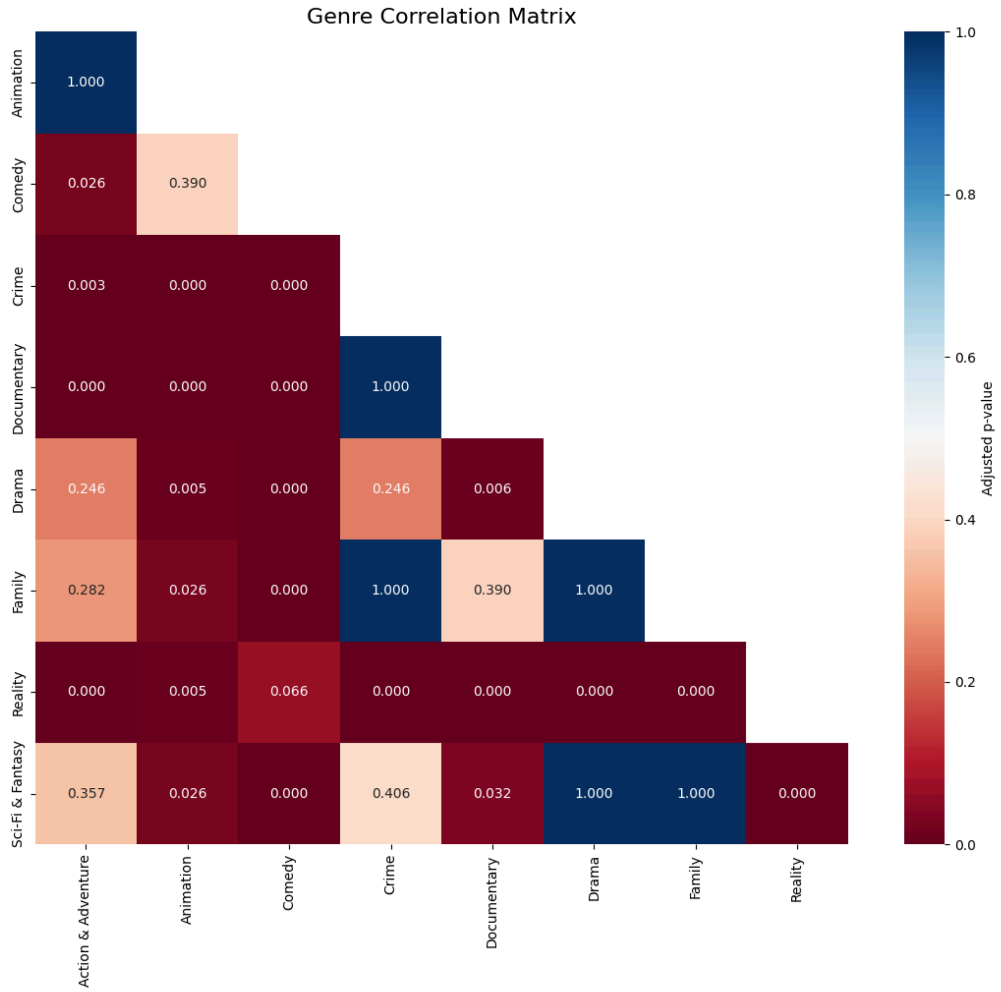

Figure 5: Genre Correlation Matrix from Pairwise Wilcoxon Rank Sum Test

C. Kolmogorov-Smirnov Test

A Kolmogorov-Smirnov test is also conducted to test for pairwise differences between genres using significance level α = 0.05.

H_0 : The distributions of the pairs of genres are statistically similar to each other.

H_1 : The distributions of the pairs of genres are statistically different from each other.

A genre correlation matrix is also made which is a representation of the various p-values and a good visualization of the differences between the genre pairs. Blue cells are correlations with a higher p-value and Red cells are correlation with a lower p-value.

Pairwise tests with p-values less than 0.05 result in the rejection of H_0, so the ratings distributions of these pairs of genres are statistically different from each other. Pairwise tests with p-values greater than 0.05 result in H_0 being unable to be rejected, so the ratings distributions of these pairs of genres are statistically similar to each other. In Fig. 7 below, the dark red cells, with p-values less than 0.05, represent pairs of genres whose distributions are statistically different from each other.

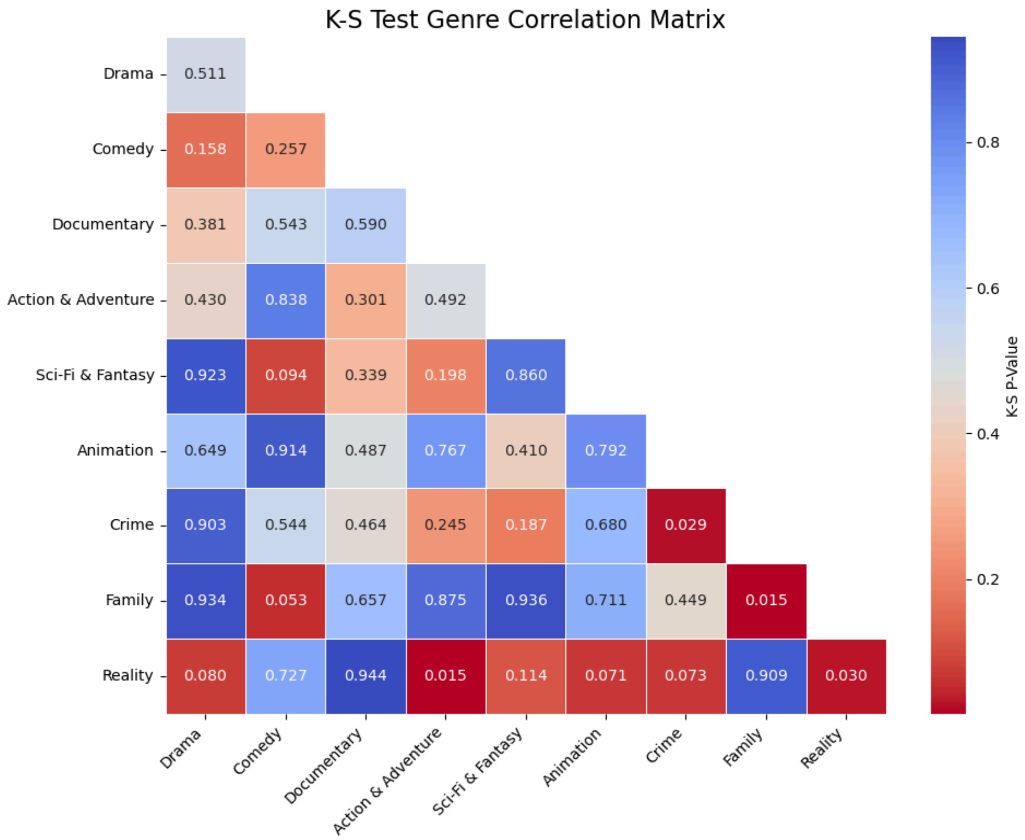

Figure 6: Correlation Matrix of the K-S P-Values across Different Genres

D. Conover post-hoc test

Finally, a Conover post-hoc test was conducted in order to explore the statistical differences between TV show ratings across various different genres. The null & alternative hypotheses remain the same as the ones in the K-S test, as well as the significance level.

Pairwise tests with p-values less than 0.05 result in the rejection of H_0, so the ratings distributions of these pairs of genres are statistically different from each other. Pairwise tests with p-values greater than 0.05 result in H_0 being unable to be rejected, so the ratings distributions of these pairs of genres are statistically similar to each other. In Fig. 7 below, the dark red cells, with adjusted p-values less than 0.05, represent pairs of genres whose distribution means are statistically different from each other.

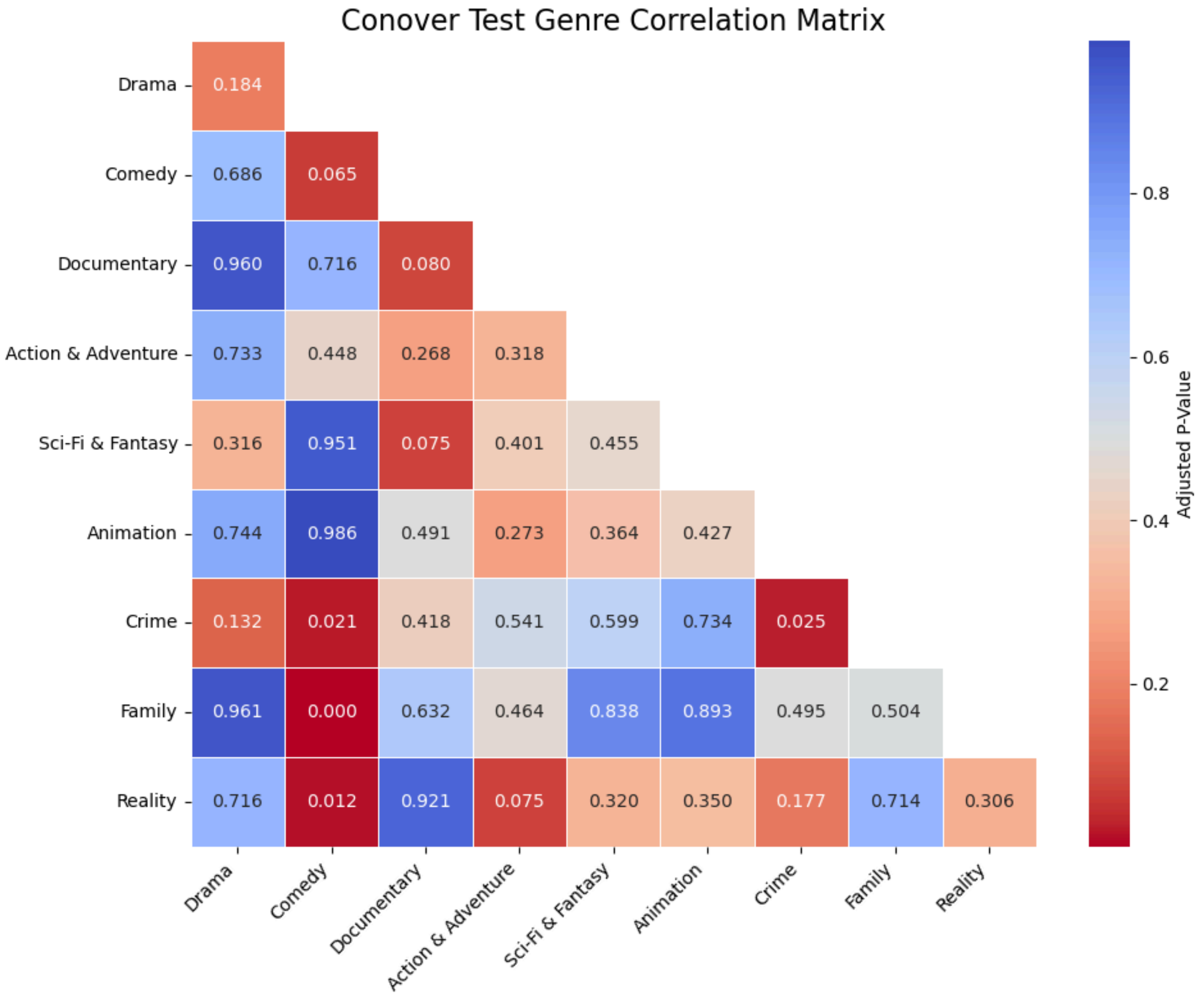

Figure 7: Correlation Matrix of the Conover P-Values across Different Genres

VII. Study Implications

In comparing the parametric Tukey test and nonparametric pairwise Wilcoxon test, the parametric test determines the pair of Reality and Comedy to have different distributions with $p_{adj} = 0.003$, but *not* in the nonparametric test. Conversely, in the nonparametric test, the pairs of Drama and Documentary and Sci-Fi & Fantasy and Documentary were determined to have different distributions with $p_{adj} = 0.006$ and 0.032, respectively, but *not* in the parametric test. Aside from these three pairs of genres, the Tukey and Wilcoxon tests agreed on the other 33 pairs.

These differences in which pairs of genres are statistically different, and also the p-values themselves, can be attributed to a variety of causes. Firstly, as seen previously in the individual Anderson-Darling tests, the vote_average data for each genre are not normal at all, meaning ANOVA tests are not appropriate for the data. Also, the data are not independent; some shows have multiple genres, and therefore appear in multiple groups, violating assumptions. Furthermore, ANOVA and Tukey pairwise comparisons test for differences in distribution means, while Kruskal-Wallis tests for differences in distribution medians, and Kolmogorov-Smirnov and Conover test for differences in overall distribution.

Differences might also be due to the parametric matrix test emphasizing linear relationships and being limited by strict assumptions about rating distributions, resulting in weaker correlations and fewer significant relationships. The nonparametric tests are more flexible, thus leading to stronger correlations and more significant and precise p-values. Given that TV show ratings often deviate from normality and may not adhere to linear trends, the nonparametric results may provide a more accurate and comprehensive view of genre relationships in the dataset.

VIII. Subject Matter Implications

The findings from this study have significant implications for the television industry, particularly in understanding audience engagement and guiding content development strategies. By demonstrating the variability in episode ratings across genres, this research highlights the importance of tailoring programming decisions to align with audience preferences. Genres such as Drama and Comedy, which consistently show distinct rating patterns, may offer opportunities for content creators to invest in high-engagement programming. Additionally, understanding underperforming genres can help producers innovate and experiment with formats to attract viewers.

From a strategic perspective, the insights gained from this study can influence how production studios allocate resources. For example, knowing which genres resonate most with audiences can inform budget prioritization, marketing campaigns, and scheduling decisions for stakeholders across the Television Industry.

IX. Further/Future Insights

While this study provided valuable insights into the relationship between TV show genres and episode ratings, it also raised several points for potential further investigation. Among these, the following three stood out as particularly significant throughout our analysis:

1) Additional Variables:
    - This study focused on episode ratings as the primary metric of audience reception, but other factors, such as show data (e.g., episode length, number of seasons/episodes), production data (e.g., budget, actor popularity, marketing strategies), viewership data (e.g., age and socioeconomic data), could provide a more comprehensive understanding of audience behavior.
2) Multigenre Interactions
    - This study treated genres as independent categories, but TV shows inherently span multiple genres (e.g., Sci-Fi & Fantasy with Drama). Future research could explore how multi-genre combinations influence ratings, identifying whether certain combinations perform better than others and why.
3) Temporal Trends

- This study analyzed TV shows within this past calendar year (2024), but audience preferences constantly evolve over time. Examining how ratings for different genres change over time may offer deeper insights into not just genre ratings, but also broader trends in TV show consumption and audience engagement.

Appendix

1) Data Cleaning:
    - Data Cleaning was conducted using Python, and the relevant code and results are documented in the first Python file attached in the appendix.
2) Exploratory Data Analysis, ANOVA, Tukey, Kruskal-Wallis, Wilcoxon
    - Nonparametric statistical tests, including the Kruskal-Wallis test and Wilcoxon Pairwise Ranked Sum Test, were performed in R. Details of the analysis are included in the second file attached in the appendix.
3) Visualizations for Tukey and Wilcoxon
    - The pairwise comparison results from the Tukey and Wilcoxon tests (which were performed in R) were transferred to Python to create easily understandable visualizations, and the code is included in the third file attached in the appendix.
4) Kolmogrov-Smirnov, Conover
    - Additional nonparametric tests, including the Kolmogorov-Smirnov test and Conover post hoc test, were implemented in Python. The corresponding Jupyter Notebook is provided as the last file in the appendix.