

# Non-Parametric Project

## Analysis of Rotten Tomatoes Movie Data

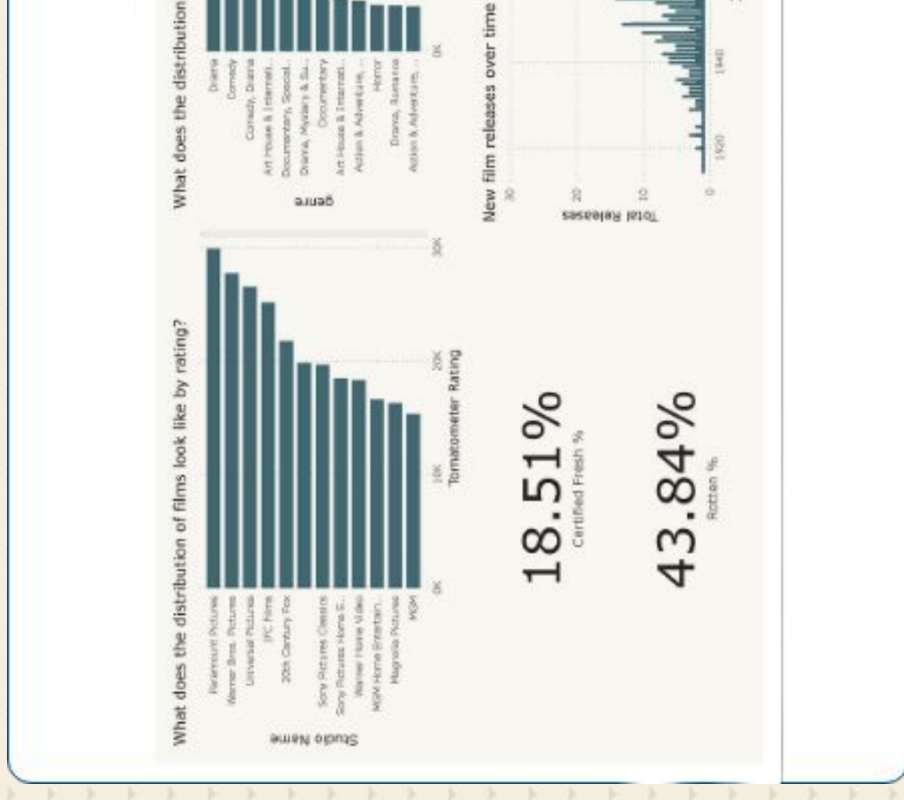
Anusua Paul – 2401800056 | Shreya Saha – 2410800095

Sunanda Maity – 241080102 | Varsha Gupta – 241080107

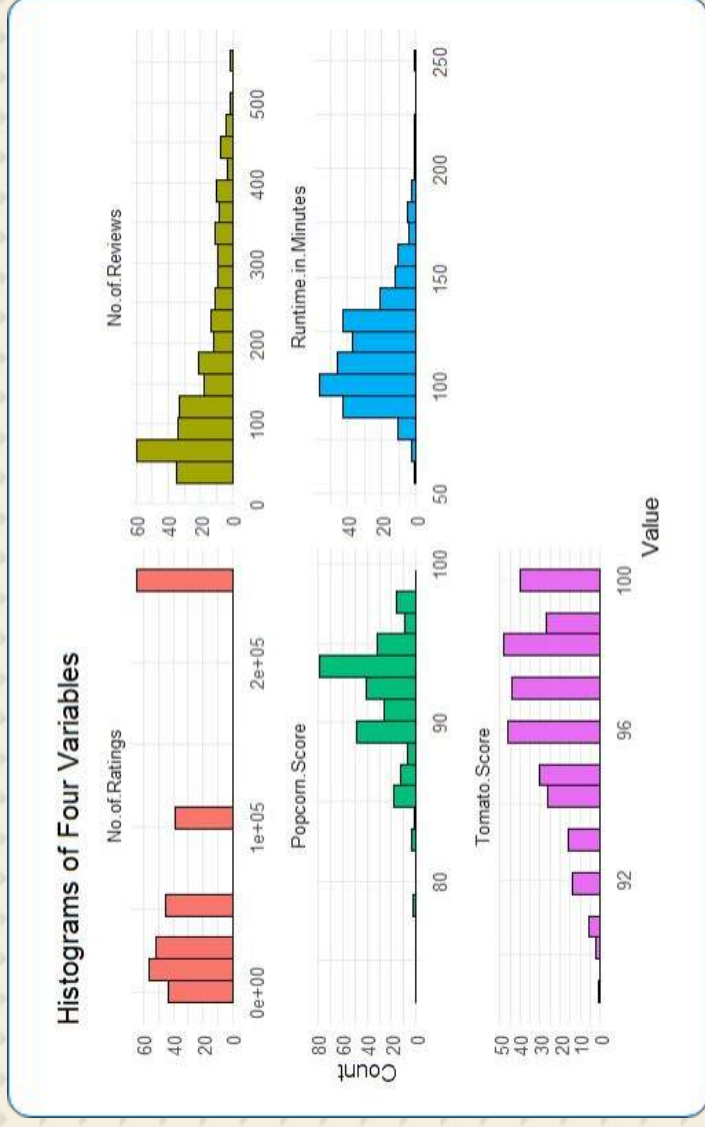
**Group 2**

# Introduction & Dataset

- ▶ **Objective:** To apply non-parametric statistical tests to understand relationships within movie data.
- ▶ **Data Source:** A curated dataset of the Top 300 movies from Rotten Tomatoes.
- ▶ **Key Variables:**
  - ▶ Tomato Score, Popcorn Score
  - ▶ No. of Reviews, No. of Ratings
  - ▶ Runtime, Genre, Release Year
- ▶ **Comment:** The goal is to use tests that don't require our data to be normally distributed (which it isn't).

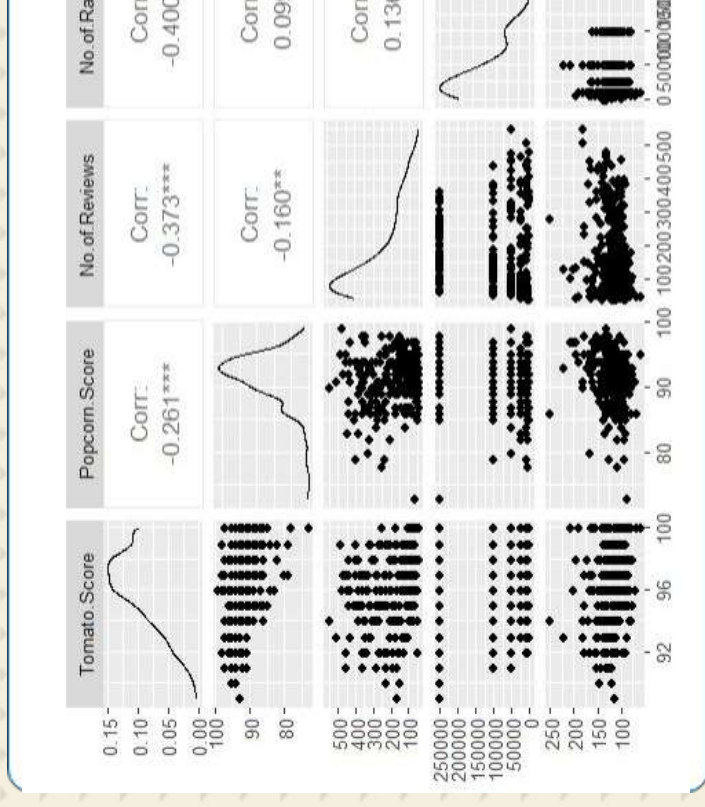


# Exploratory Data Analysis (EDA)



Histograms show variables have different distributions. Tomato

Score is left-skewed; Reviews/Ratings are right-skewed.



The `ggpairs` plot gives a first look at relationships

correlations present among the fe

# Analysis 1: Is the Data Random?



## Test & Why

**Test:** The Runs Test

**Why?** We need to know if our data is a random sample before applying other tests.



## Hypothesis

**Null (H0):** The data is random.

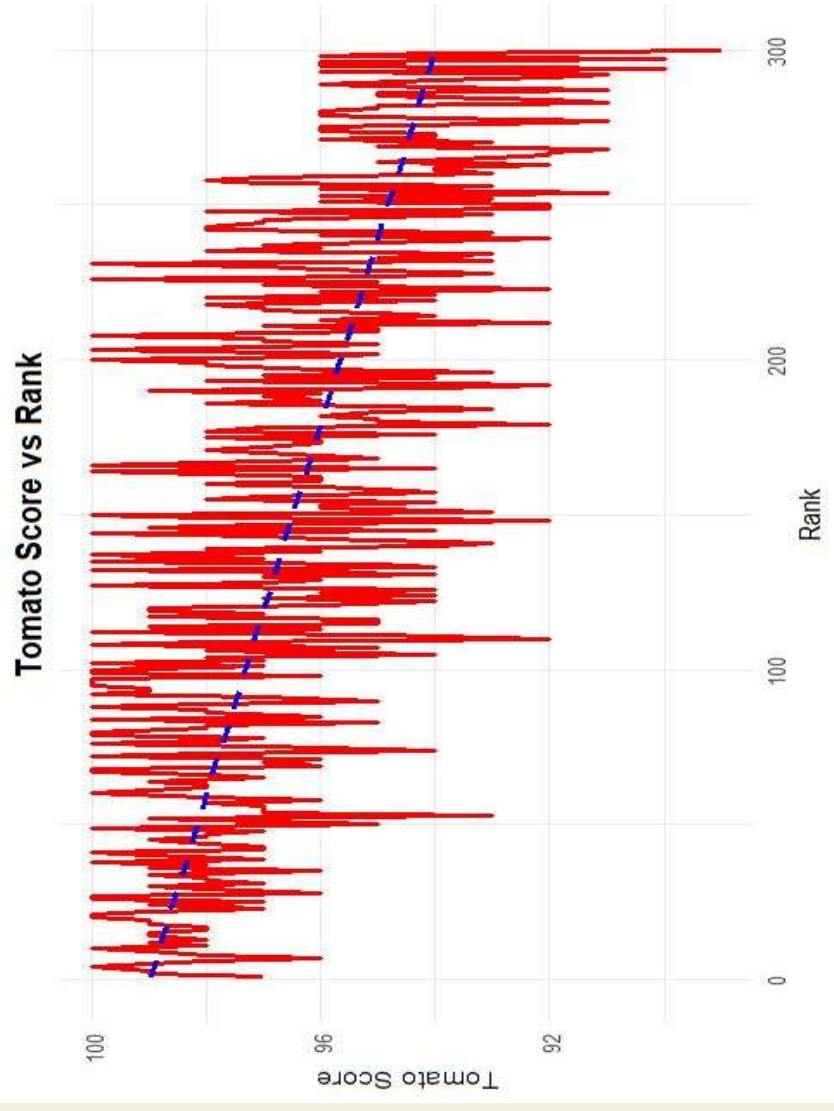
**Alt (H1):** The data is NOT random.

## Conclusion

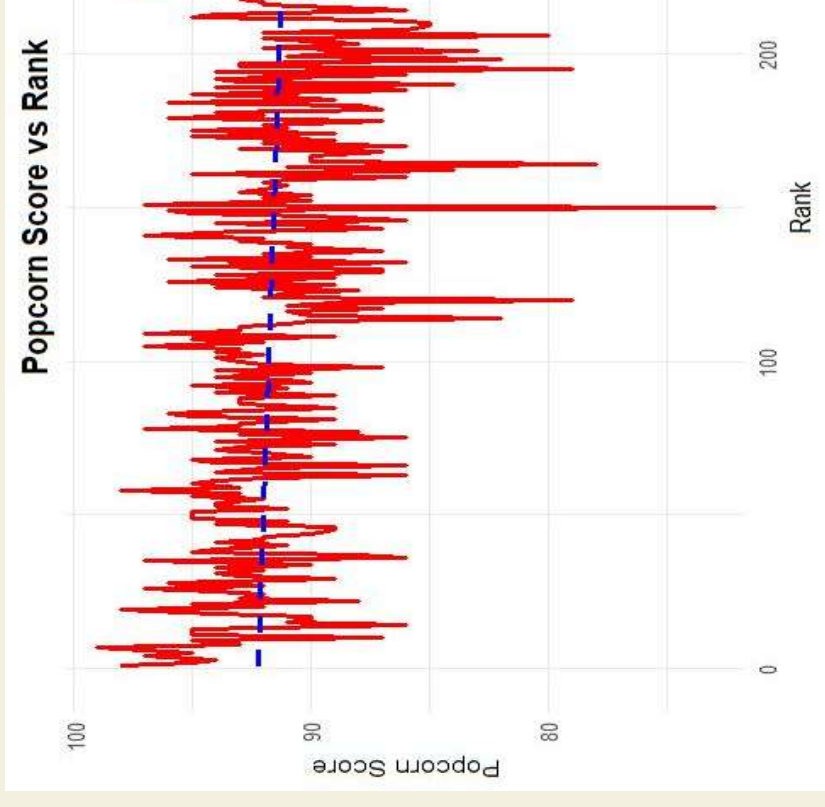
**Tomato Score:** Can data be random (Reject H0)?

**Others:** Random (Fail to reject H0)

**Comment:** Tomato data is random because the data is sorted by rank.



This plot shows a clear negative relationship between Tomato Score and Rank. As rank increases, Tomato Score decreases almost linearly, indicating that movie rankings are primarily based on the critic's Tomato Score.



In contrast to Tomato Scores, Popcorn Scores show a flat pattern with Rank. The trend line is nearly flat, indicating that Popcorn Scores are not determined by audience scores — they're based on critic scores. We may expect same plot for other

# Analysis 3: Are Reviews and Ratings Related?

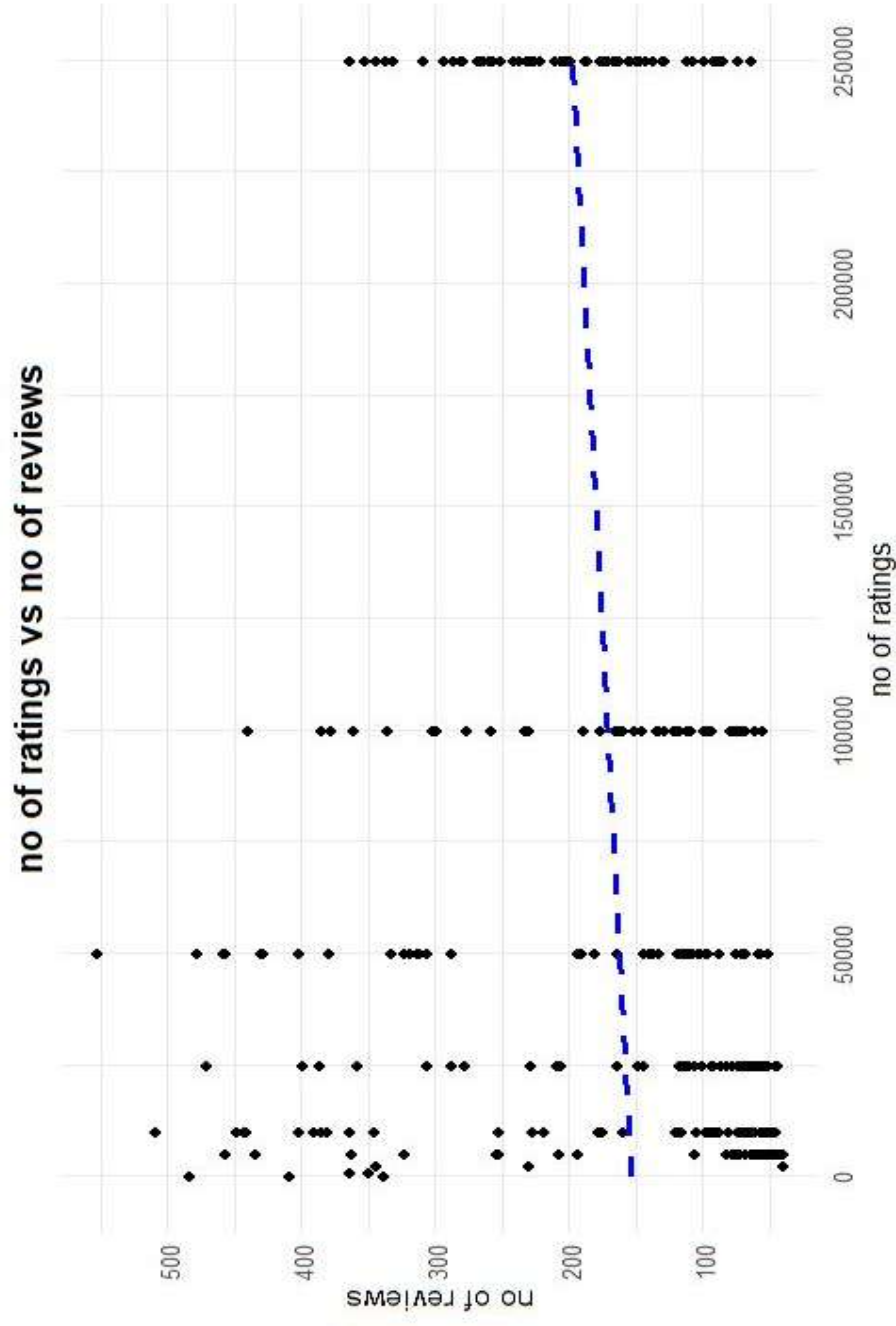
## Spearman's Rank Correlation

- ▶ **H0:** There is no correlation.
- ▶ **Result:** Very small p-value ( $p = 1.428e-8$ ).
- ▶ **Decision:** Reject H0.

## Kendall's Tau

- ▶ **H0:** two features are independent.
- ▶ **Result:** Very small p-value ( $p = 1.344e-$
- ▶ **Decision:** Reject H0.

**Conclusion:** Yes, there is a strong, significant positive correlation. As the number of critic reviews increases, audience ratings also tends to increase.



### Interpretation:

This plot shows a weak positive relationship between the number of critic reviews and audience ratings. Movies that receive more reviews generally also get higher ratings, though the correlation is not very strong.

# Analysis 4: Does Runtime Affect Engagement?

## Q1: Runtime vs. No. of Reviews

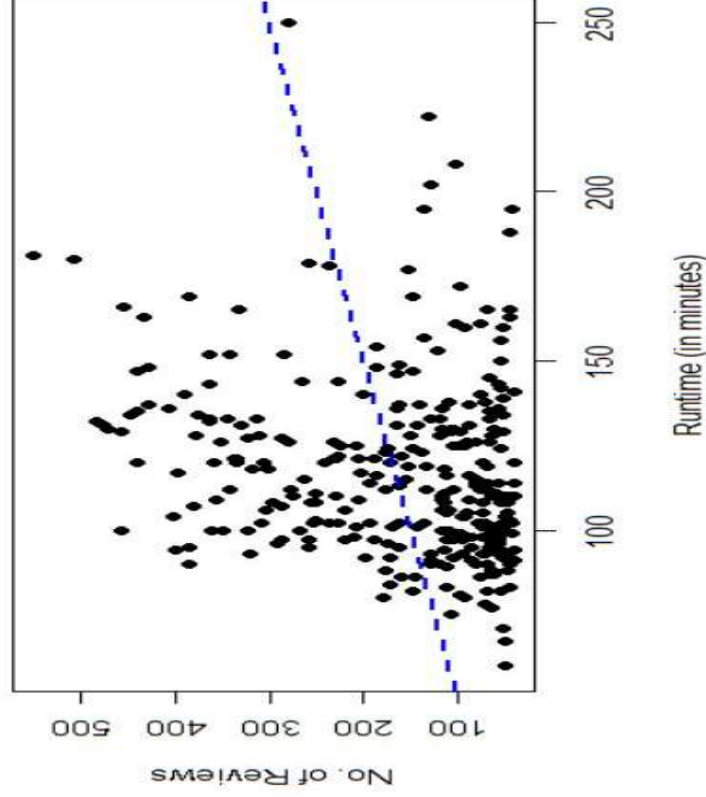
- ▶ **Test:** Spearman/Kendall's Correlation.
- ▶ **Result:** p-value  $< 0.05$  in both case. Reject  $H_0$ .
- ▶ **Interpretation:** Yes, a significant positive correlation. Also they are not independent. Longer movies tend to have more reviews.

## Q2: Runtime vs. No. of Reviews

- ▶ **Test:** Spearman/Kendall's Correlation.
- ▶ **Result:** p-value  $> 0.05$ . Can not reject  $H_0$ .
- ▶ **Interpretation:** No significant correlation. They are independent. Thus movie length does not affect how many rating the movie gets.

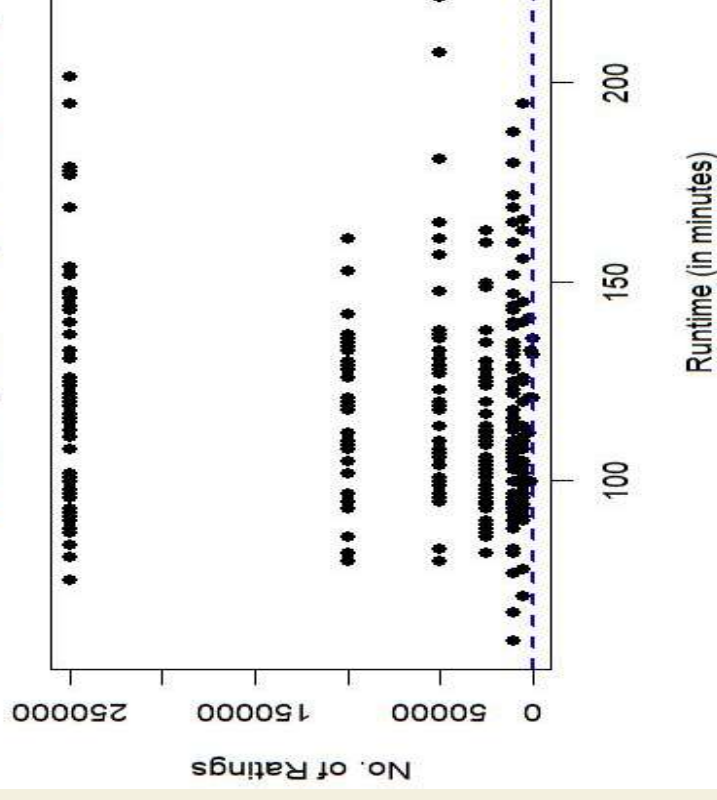
**Comment:** Longer movies tend to get more number of reviews, whereas, there is no effect of length of movies on rating.

Runtime (in minutes) vs No. of Reviews



This plot shows a weak but significant positive relationship between runtime and number of reviews. As runtime increases, the number of critic reviews also tends to increase i.e longer movies generally attract more reviews.

Runtime (in minutes) vs No. of Ratings



The scatterplot is showing a flat trend with no visible relationship between runtime and number of ratings. This implies that the number of audience ratings is not systematically influenced by runtime. In other words, **runtime does not influence how many people rate a movie.**

# Analysis 5: Do Tomato Scores Differ by Genre?

**Question:** Do different genres have different median

Tomato Scores?

**Test Used:** Kruskal-Wallis Test

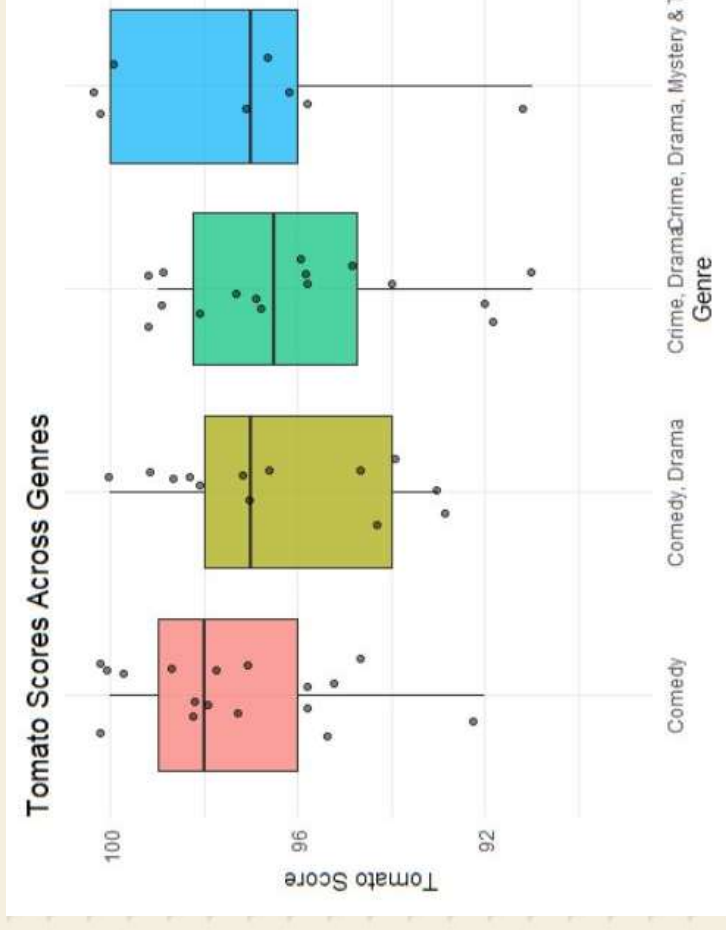
**Why?** A non-parametric alternative to ANOVA for comparing 3+ groups.

**Null Hypothesis (H<sub>0</sub>):** The median Tomato Score is the same across all genres.

**Interpretation:** With a p-value = 0.2897 > 0.05, we do not reject H<sub>0</sub>.

This means — there's no significant difference in Tomato

Scores across genres. In other words, critics rate different genres similarly.



# Analysis 6: Have Scores Changed Over Time?

**Question:** Are Tomato Scores for movies before 2000 different from those after 2000?

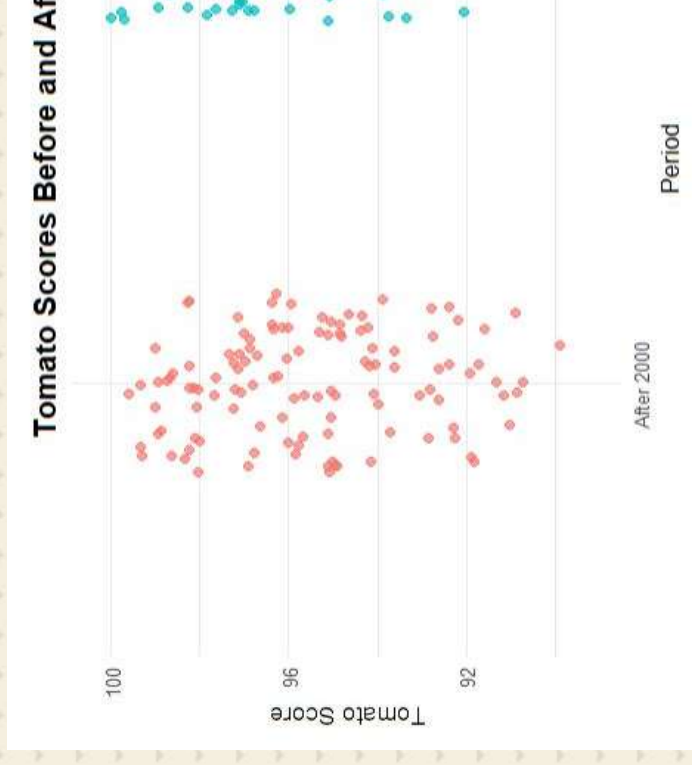
**Test Used:** Wilcoxon Rank-Sum Test (Mann-Whitney U)

**Why?** This test compares the medians of two independent groups.

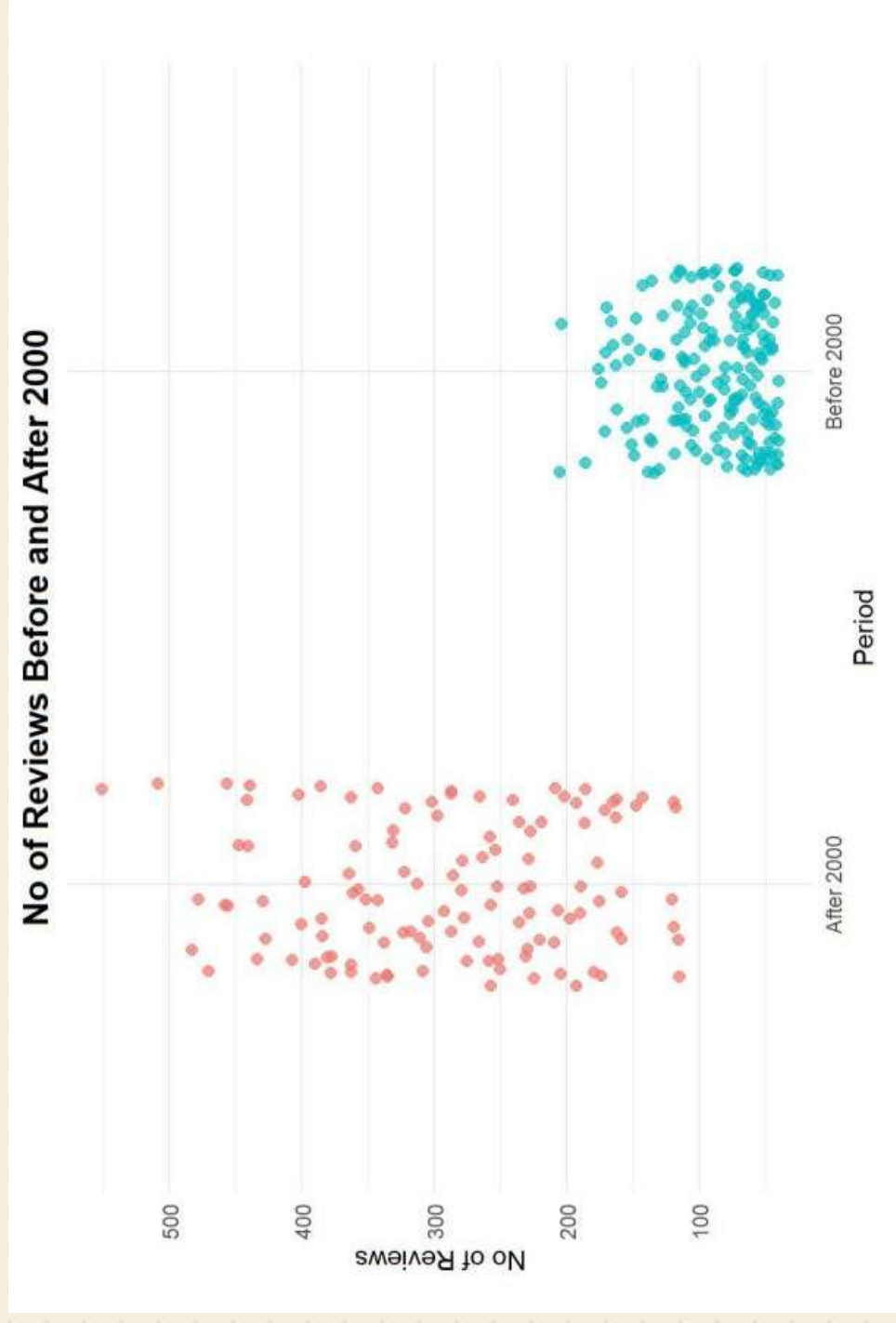
**Null Hypothesis ( $H_0$ ):** The median scores are the same for both periods.

**Interpretation:** With a p-value  $< 0.05$ , we reject  $H_0$ .

This indicates there is a significant difference in Tomato scores between older and newer films.







# Analysis 7: Have number of reviews Changed Over



## Interpretation:

From this plot, it is clear that the number of reviews has decreased for the movies released before 2000. So it means that for less number of reviews, these movies got less rank and led to high tomato rank and better rank.

# Final Conclusions

-  **Ranks:** Movie ranks in this dataset are driven by Critic Scores (Tomato Score), not Audience Score (Popcorn Score).
-  **Engagement:** No. of Reviews and No. of Ratings are strongly correlated. Runtime is positively correlated with no. of reviews, suggesting longer movies attract more critical attention. Whereas, there is no significant correlation between runtime and no. of ratings, implying rating does not depend on movie length.
-  **Group Differences:** Critic scores do not differ significantly across genres. But we found a significant difference in tomato scores for movies made before and after the year 2000. And also we get that that number of reviews for the movies before 2000 are few than the movies after 2000.
-  **Overall:** Non-parametric tests allowed us to find significant relationships in our data without assuming a normal distribution, revealing key insights into movie rankings and engagement.

THANK YOU