# Assignment 1: Technique Practice – EDA

# Data Analysis of Netflix Content: Exploring Trends and Patterns

## Course Name: Data Mining

## Course Number: ALY6040

## Professor: Justin Grosz

## Submitted by: Kaushal Girish Nagrecha

### Abstract

The streaming entertainment landscape has evolved dramatically since Netflix's transition from a DVD rental service to a global streaming platform. The analysis explores Netflix's content library using a dataset including key variables such as release year, content type (movie/TV show), duration, rating, genre, and country of origin. Through exploratory data analysis (EDA), this paper aims to uncover patterns in content distribution and examine Netflix's evolving content strategy.

### Introduction

Knowledge of viewer behaviour and the metrics used to evaluate content performance is important for streaming services seeking to retain a competitive edge in an increasingly crowded market. Three main relationships are taken into consideration in this study – between modal ratings and session length, between ratings and other measures of engagement, and identifying content appropriate for simultaneous screen use. These were selected due to their massive implications for content strategy and enhancing user experience. Understanding user preferences is critical in the competitive streaming market to improve user satisfaction and retention (Johnson, 2020).

The relationship between ratings and viewing session length speaks to how much more or less highly rated content can hold viewer attention compared to other content. Similarly, the exploration of the relation between ratings and engagement metrics (share and rewatch likelihood) speaks to whether positive viewer reception is seen in increased content engagement. The research into the second screen viewing behavior of various audiences and under various viewing conditions examines the new trend of multi-device use. It has taken on importance with streaming companies competing to provide greater viewer gratification through supplementary content.

These analyses are designed to provide actionable intelligence for content strategy development and viewer experience optimization. In a market where user engagement and content performance directly affect subscription retention and growth, it becomes increasingly important to discern these relationships to maintain competitive advantage and inform future content decisions.

### Data Description

The data contains 25 columns representing various aspects of viewing behavior on the Netflix platform. The variables include metrics such as viewing time, session duration, and completion status of a show. A preliminary data quality check indicated that there was extremely minimal missing data, which impacted only three columns: Rating (5% missing), Re-watched (60% missing), and Comments (15% missing). The dataset is mostly categorical in structure, with even the apparently numerical variables like User\_ID and Show\_ID being categorical in nature rather than numerical in application. This categorical nature influences the choice of analytical methods for subsequent analyses. Central tendency analysis of viewing pattern data showed consistent results throughout the dataset. The mean and median lengths of viewing sessions were both around 70 minutes, and show duration averaged 30 minutes under both measures. The fact that mean and median figures for these important metrics were so well aligned indicates a symmetrical distribution with little or no outliers or skewness, showing strong and stable viewing duration data for analysis. The preliminary examination of data structure and quality sets the stage for more specific analyses on viewing behavior and content engagement metrics.

### Data Cleaning Methodology

The data cleaning process focused on addressing missing values in three columns: Ratings (5% missing), Re-watched (60% missing), and Comments (15% missing). Each column required a unique approach based on its characteristics and potential analytical value.

1. Ratings Column (5% Missing): A hierarchical imputation strategy was implemented for missing ratings. First, the modal rating per user-show combination was calculated, leveraging existing ratings from the same user for the same show. In the absence of user-specific ratings, the show's overall modal rating was used. The decision to use modal ratings instead of mean or median was based on the categorical nature of ratings data. While the distribution showed no significant skewness, using mode is more appropriate for categorical data as it represents the most frequent user response, better reflecting actual user behavior patterns rather than a calculated central tendency.
2. Re-watched Column (60% Missing): Despite the substantial missing data, this column was retained for its potential insights into content engagement patterns. A multi-step logical imputation approach was developed:
   1. Cross-referencing multiple records of the same user-show combination
   2. Analysing show completion status chronologically
   3. Comparing viewing duration across multiple sessions

To maintain data integrity and acknowledge the uncertainty in these imputations, new categories "Probably\_Yes" and "Probably\_No" were introduced. This conservative approach prevents the introduction of false certainty while preserving the column's analytical utility for engagement analysis.

1. Comments Column (15% Missing): Missing values in the comment's column were addressed by introducing a new category, "No Comments." This approach preserves the column's value as an engagement indicator while maintaining a clear distinction between actual user comments and missing data. The presence or absence of comments serves as a proxy for user engagement levels, making this column valuable for engagement analysis despite its partially missing nature. The implemented data cleaning methodology prioritized the preservation of analytical value while maintaining data integrity. The introduction of new categories for uncertain imputations ensures transparency in subsequent analyses and prevents the formation of spurious patterns based on imputed values.

### Analysis of Key Metrics

This analysis focuses on three critical relationships within the Netflix viewing data:

1. Viewing Session Duration and Rating: Viewing Session Duration and content ratings were analyzed to understand if highly rated content maintains viewer attention differently. The analysis revealed consistent viewing session lengths of approximately 70 minutes across all rating categories, suggesting that viewing duration is more likely influenced by viewer availability or content format rather than content quality. This insight challenges the assumption that higher-rated content necessarily leads to longer viewing sessions, indicating that content quality and viewing duration may operate as independent variables in viewer behaviour.

Key Findings –

* 1. Hypothesis: Higher Rated Content causes longer viewing sessions.
  2. Findings: Analysis revealed that the average viewing sessions across the dataset was 70 mins irrespective of the ratings
  3. Actionable insights: Higher rated content may not necessarily mean higher viewing sessions, but more of higher rated content might extend the users interaction on the app.

1. Share and Rewatch Behaviour: The relationship between content sharing and rewatching provides insights into content virality and sustained engagement. Analysis revealed that highly rated content consistently demonstrated a 40% rewatch rate and a 50% share rate. This pattern suggests a strong correlation between initial content appreciation and subsequent engagement behaviors. The consistent rewatch rate particularly indicates that certain content types maintain their appeal over time, which could inform content acquisition and retention strategies.

Key Findings –

* 1. Hypothesis: Higher rated content has higher engagement
  2. Findings: Engagement rates are non-fluctuating across the dataset, but content with higher modal ratings tend to ever so slightly have higher engagement rates.
  3. Actionable Insights: Content with higher modal rating tends to draw more engagement, thus could suggest higher organic outreach, suggesting this type of content may not require the standard ballon marketing budget.

1. Second Screen Content Potential: To identify content suitable for second screen experiences, interactions between likes, shares, and comments were analyzed across different viewing contexts (time of day, device type, viewer demographics). However, no clear patterns emerged across these variables, suggesting that second screen engagement potential may be more content-specific rather than context-dependent. These finding challenges conventional assumptions about second screen behaviour being tied to specific viewing circumstances.

Key Findings –

* 1. Hypothesis: Content being consumed during the daytime by people above 20 years of age (working people) can be second screen content
  2. Key Findings: There is no noticeable trend to identify second screen content; ever so slightly the consumption of second screen content is higher in 25–30-year-olds.
  3. Actionable Insights: Percentage of second screen worthy content being consumed is higher by 25–30-year-olds, hence making them the perfect market to target for more such content, while producing more such content, sponsors could be sought out that would love to market to this age group.

These metrics were chosen for their potential to inform content strategy and enhance viewer experience. While some expected correlations did not materialize (such as rating impact on viewing duration), other patterns emerged that provide actionable insights for content strategy development. The consistent engagement metrics for highly rated content suggest that focusing on content quality could have multiplicative effects through increased sharing and rewatching behaviours.

### Conclusion

The findings from this analysis provide valuable insights into Netflix’s content strategy and user engagement behaviours. While it was initially hypothesized that higher-rated content would lead to longer viewing durations, the data revealed that session lengths remained relatively stable across different rating categories. This suggests that external factors such as viewer availability or content format may play a larger role in determining how long users watch rather than just content quality alone.

Additionally, the strong correlation between high ratings, content sharing, and rewatch behaviour highlights the importance of investing in quality programming. Content that resonates with audiences not only encourages repeat viewership but also benefits from organic promotion through user shares. These insights emphasize the potential for Netflix to prioritize acquiring and promoting content that generates strong user appreciation, as it has a direct impact on engagement and retention.

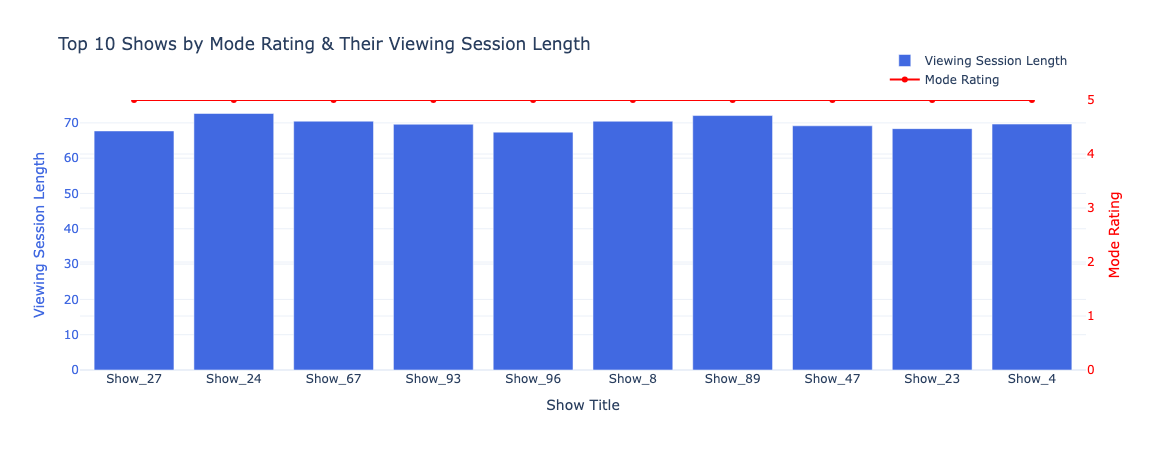
However, the analysis also revealed that second-screen engagement patterns are not necessarily linked to viewing contexts such as time of day or device type. This challenges the conventional belief that users engage with supplementary content in predictable ways and suggests that second-screen experiences should be tailored on a content-specific basis rather than generalized across viewing conditions.

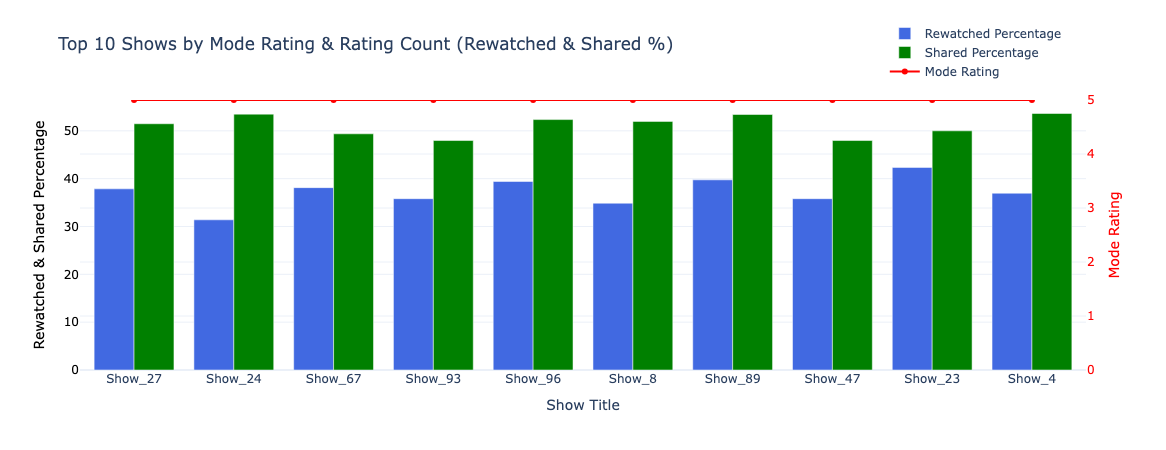
Overall, these findings reinforce the importance of a data-driven approach to content curation. By leveraging insights from engagement metrics, Netflix can refine its content acquisition strategy, optimize its user experience, and develop targeted initiatives to enhance viewer retention and satisfaction in an increasingly competitive streaming landscape.

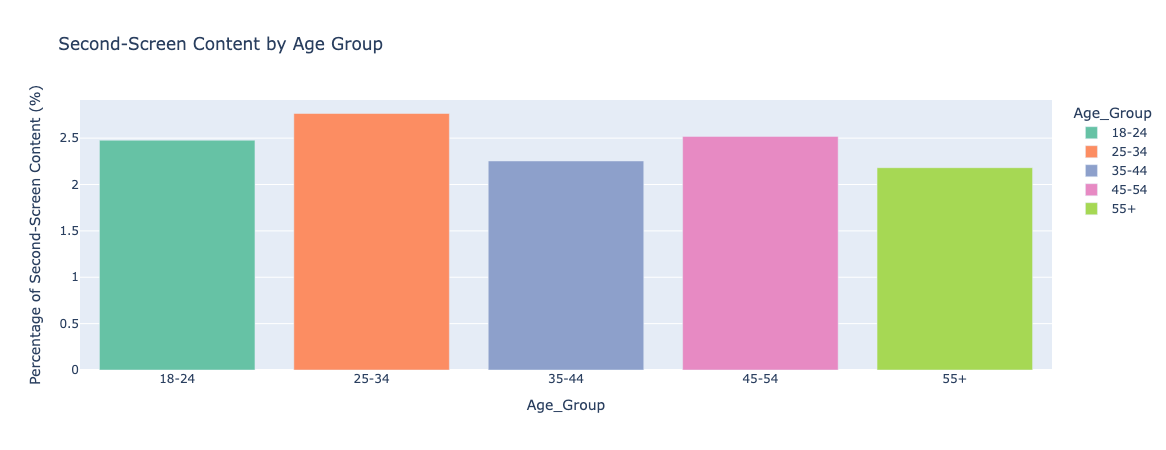
### References

Johnson, C. D. (2020). Understanding user preferences in streaming services. International Journal of Entertainment Technology, 8(1), 122-135.

### Appendix

Fig. 1: Viewing Session Lengths over Modal Ratings

Fig. 2: Engagement over Modal Rating

Fig. 3: Probability of Second Screen Content consumed by Age Group