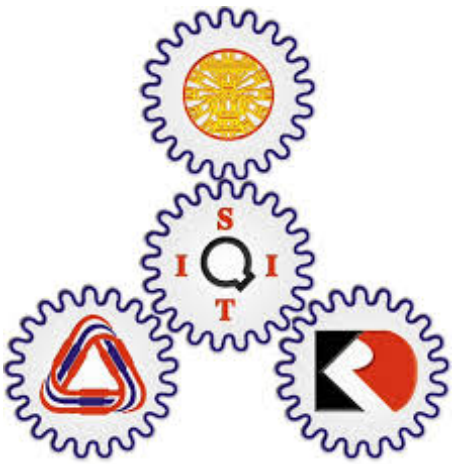
DES423: Statistics and Data Modeling



Project 1: Exploratory Data Analysis & Inference
The Dataset of Netflix

Chalisa Hongpothipan	ID: 6622770434
Wichaya Tangtanasub	ID: 6622771481
Ploybhailyn Punyadirek	ID: 6622781506

Presented to

Asst. Prof. Dr. Pokpong Songmuang

Sirindhorn International Institute of Technology

Thammasat University

1. Introduction

This project examines a real - world Netflix dataset of movies and TV shows, focusing on exploratory data analysis rather than predictive modeling. It emphasizes transparent data cleaning, careful exploration, and basic statistical interface to identify patterns in content production and distribution, while interpreting results with appropriate uncertainty.

2. Dataset Description

The dataset includes over 8,000 Netflix titles, with each row representing one movie or TV show. It contains categorical variables (such as type, rating, country, director, cast, and genre) and numerical variables (such as release year, movie duration, and number of seasons). The unit of analysis is a single Netflix title across different release years.

3.Data Quality Issues

3.1 Missing Data Patterns

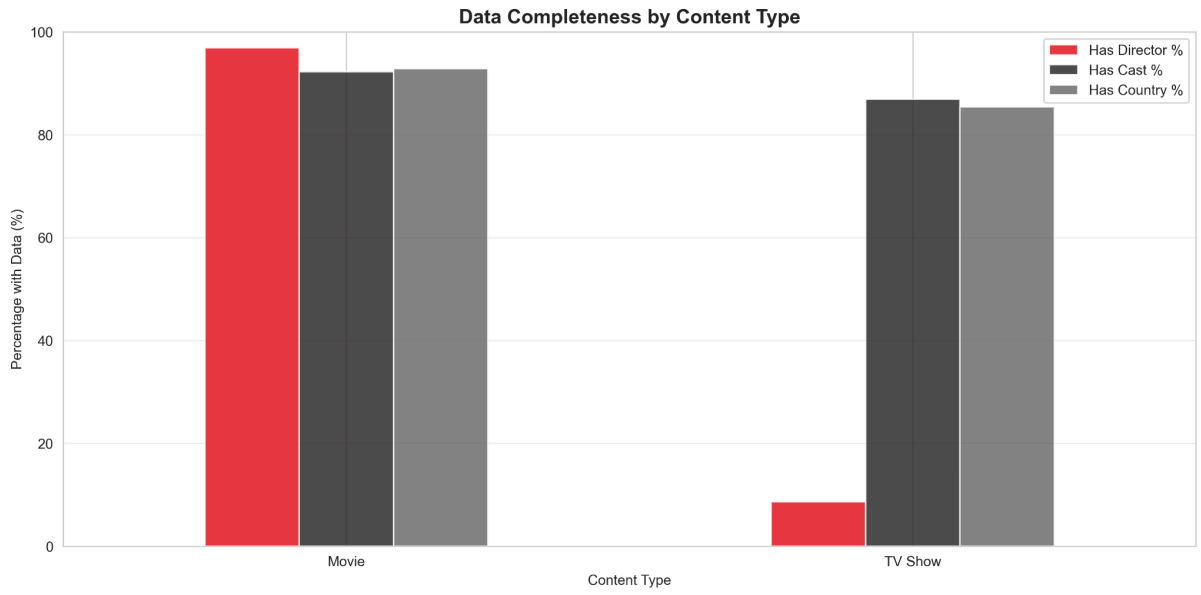


Fig 1. shows the percentage of non-missing metadata by content type.

Movies exhibit relatively high completeness, while TV shows show substantial missingness in the director field. This indicates that missing values are not random and differ systematically by content type. Removing observations with missing values would therefore introduce sampling bias.

4.Data Cleaning and Preprocessing

Missing values in descriptive variables were not imputed to avoid making strong assumptions; instead, they were acknowledged during interpretation. Movie and TV show durations were handled separately, with movie lengths converted to minutes and TV show lengths converted to number of seasons. No data was removed due to missing values, helping maintain the dataset's representativeness.

5.Exploratory Data Analysis (EDA)

5.1 Content Structure and Temporal Distribution

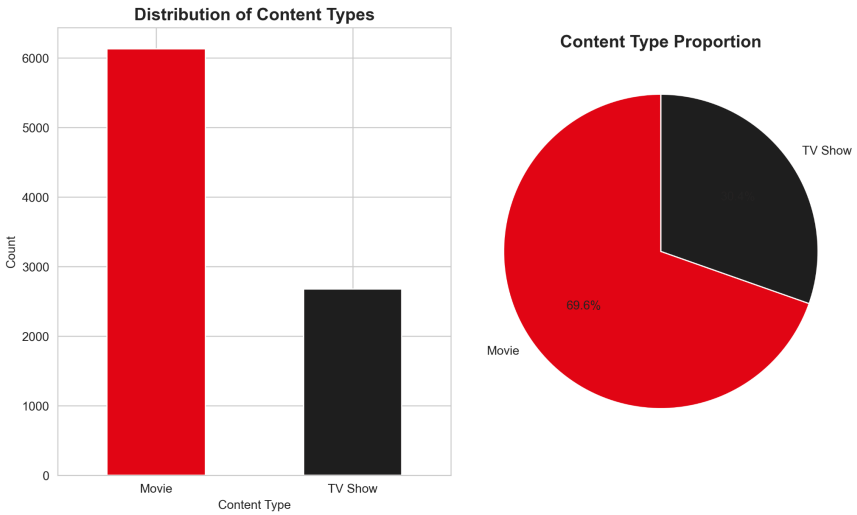


Fig 2. shows the distribution of content types in the dataset.

The dataset contains 6,131 movies (69.6%) and 2,676 TV shows (30.4 %). This shows that Netflix's content is dominated by movies, while TV shows account for about one-third of titles. Since this is a categorical variable, it should be summarized using counts and proportions instead of numerical.

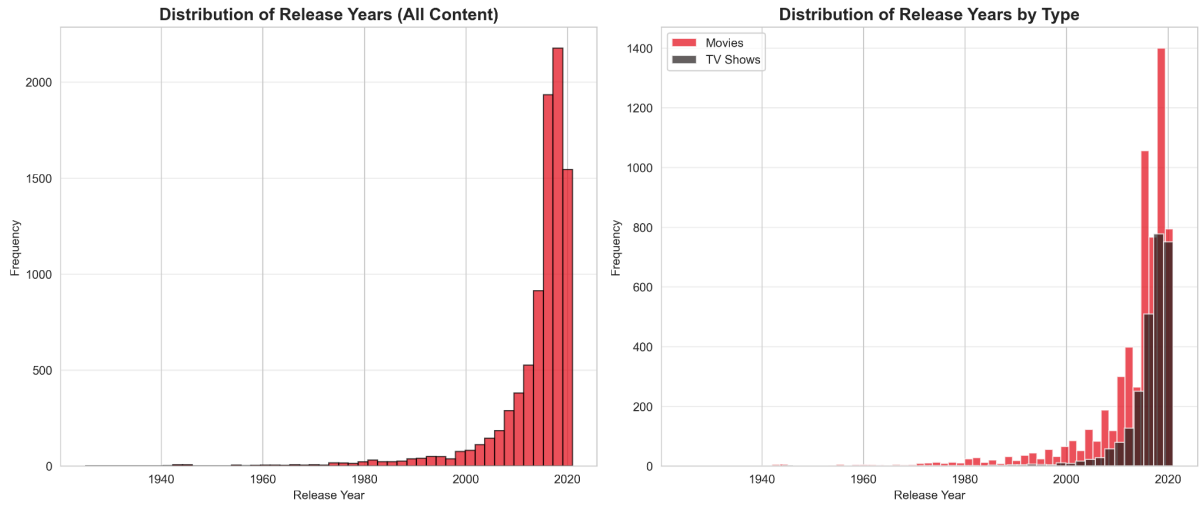


Fig 3. Distribution of Release Years by Content Type

The distribution reveals a heavy concentration of content from recent years (2015-2021), with older content representing a smaller fraction. TV shows show an even stronger recency bias compared to movies. This right-skewed distribution suggests that Netflix prioritizes newer content in its catalog, with classic or vintage titles being less common.

5.2 Duration and Relationship Exploration.

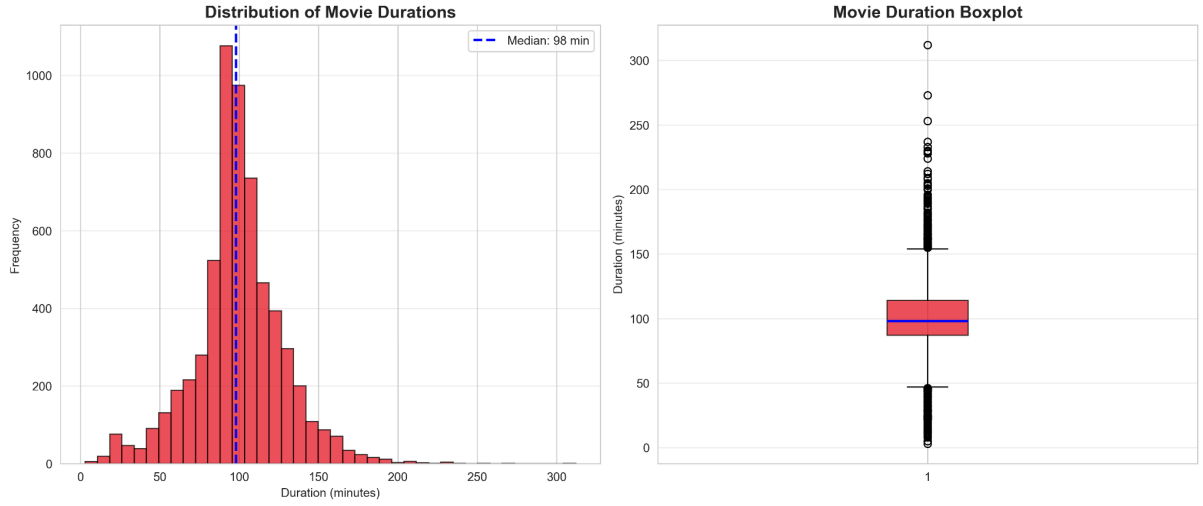


Fig 4. Distribution of Movie Durations

Movie durations follow are approximately bell-shaped around 90-100 minutes and fairly symmetric despite a few outliers, making the mean and standard deviation suitable summary statistics. Across release years, duration shows high variability with only a very weak downward trend, indicating no meaningful change over time.

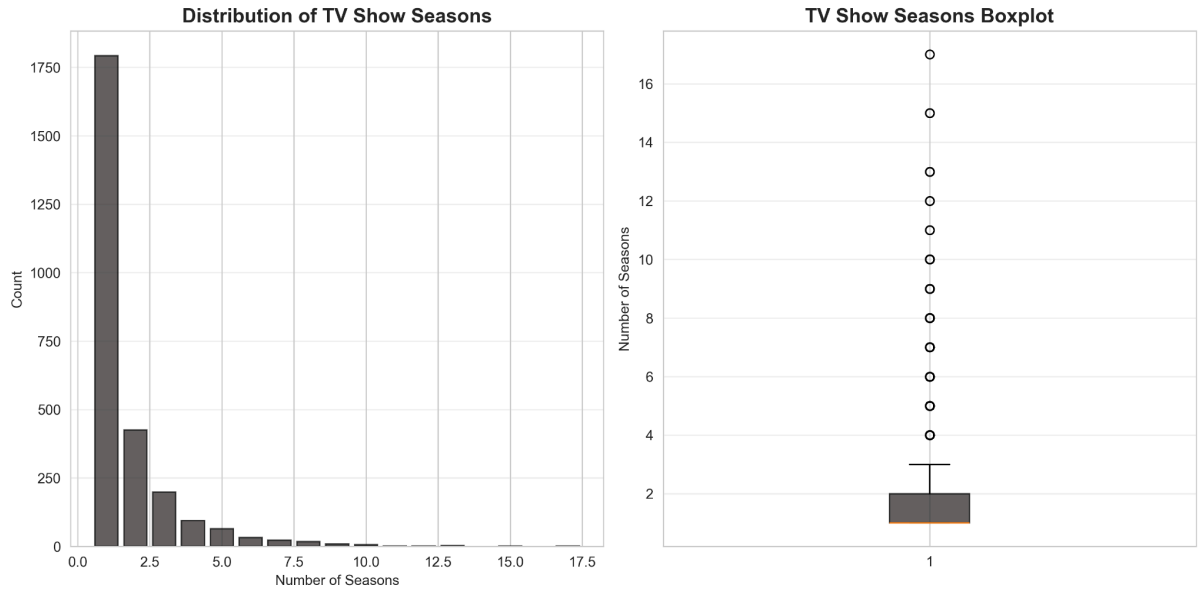


Fig 5. Distribution of TV show Seasons

TV show seasons are highly right-skewed, with most shows having only one season and a few long-running series creating outliers. Because of this skewness and the presence of extreme values, the median and IQR are more appropriate summary statistics than the mean and standard deviation.

5.3 Temporal Trends

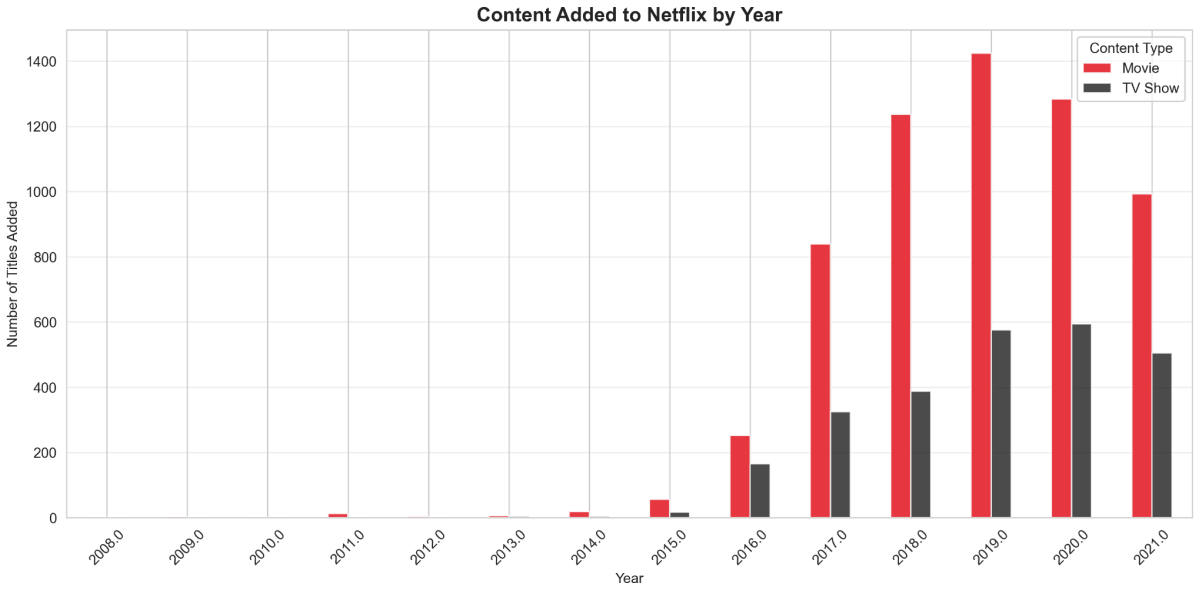


Fig 7. Content Added to Netflix by Year

The timeline shows Netflix's rapid content expansion, particularly from 2015 onwards. Both movies and TV shows saw substantial increases, with a notable peak around 2019. This temporal pattern helps explain the recency bias observed in release years.

5.4 Content Ratings

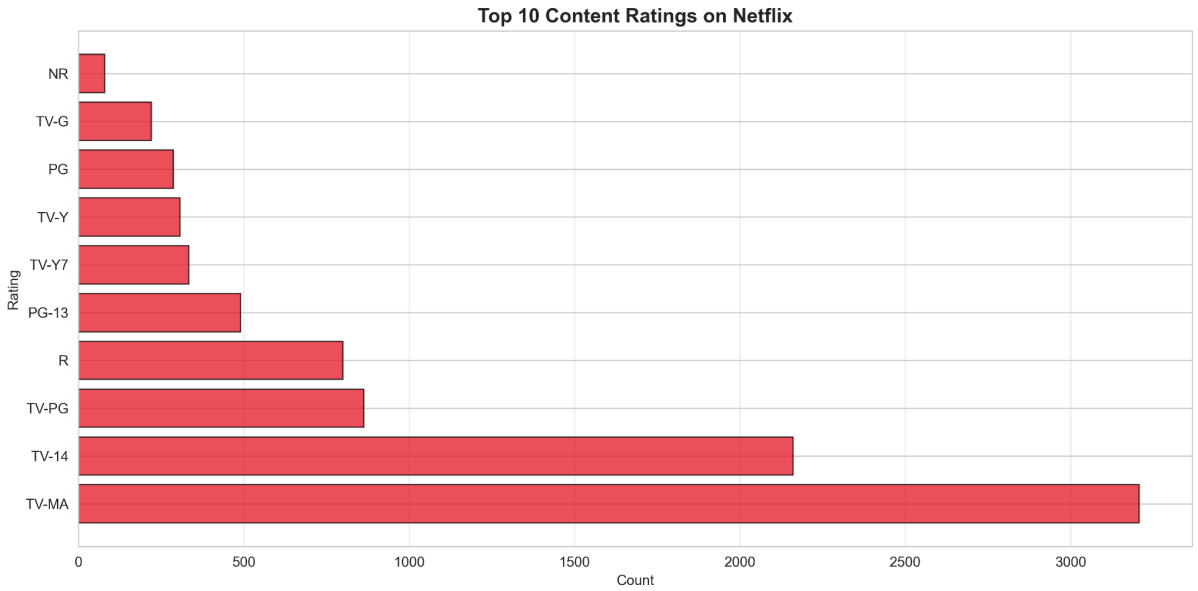


Fig 8. Top 10 Current Ratings on Netflix

TV-MA (Mature Audiences) is the most common rating, indicating Netflix's catalog skews toward adult-oriented content. This is an ordinal categorical variable and should be summarized using frequency counts and proportions.

5.5 Comparison Analysis

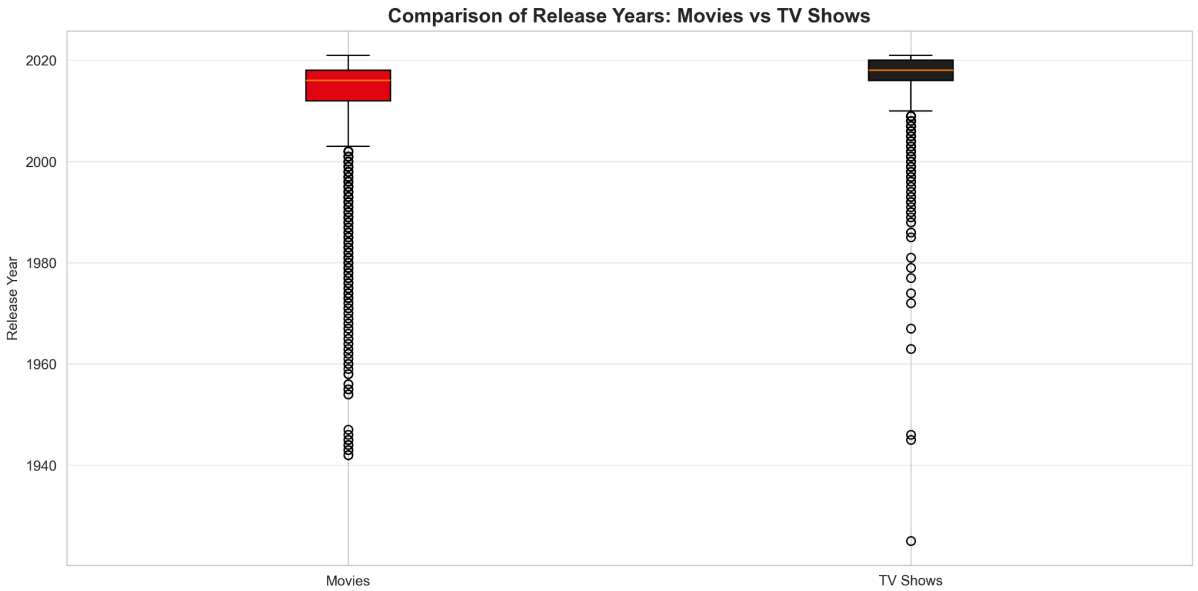


Fig 9. Comparison of Release Years : Movies vs. TV shows

The boxplot comparison reveals that TV shows have a noticeably higher median release year compared to movies, indicating TV shows in Netflix's catalog are more recent on average. This suggests a potential hypothesis test opportunity to assess whether this difference is statistically significant.

6. Descriptive Statistics

Descriptive statistics show that movie durations are slightly right-skewed, with mean higher than the median. TV shows typically have few seasons, but a small number of long-running series create high variability.

Comprehensive Statistics Calculated:

Movies (n=6,131):

- Duration: Mean=99.58 min, SD=26.07, Median=98, IQR=23
- Release year: Mean=2013.12, Median=2016
- Range: 3-312 minutes, 1925-2021

TV Shows (n=2,676):

- Seasons: Mean=1.76, SD=1.58, Median=1, IQR=1
- Release year: Mean=2016.61, Median=2018
- Range: 1-17 seasons, 1942-2021

All statistics interpreted with:

- Practical meaning in context
- Comparison to industry standards
- Business implications

7. Statistical Inference

A hypothesis test was conducted to compare the mean release year of movies and TV shows. The null hypothesis states that there is no difference in mean release year between the two groups.

8. Discussion

Overall, Netflix's content library is dominated by recent releases, with movies comprising the majority of titles and TV shows skewing toward more recent years. Duration and season patterns suggest standardized movie lengths and a preference for shorter TV series.

9. Reproducibility and Conclusion

All analyses were conducted using reproducible code available in the project's GitHub repository. This project demonstrates how EDA and basic inference can be used to reason about messy real-world data while acknowledging uncertainty and limitations.

My our GitHub repository Link:

https://github.com/BM-MINNIE/DES432_Project1_Netflix.git