

# Netlyse: A Netflix analysis with R and Quarto

Faranak Rahimi & Nils Rechberger

2026-01-01

## Abstract

Netflix has become one of the most influential global streaming platforms, offering a rapidly expanding catalogue of movies and television shows produced across a wide range of countries. As the platform continues to grow, understanding how its content catalogue evolves over time provides valuable insights into content strategy, internationalisation, and audience engagement. This project presents a data-driven analysis of Netflix's content catalogue using publicly available datasets and reproducible analytical workflows implemented in R and Quarto.

The analysis is based on two complementary datasets: a dataset containing metadata on Netflix movies and TV shows, and a dataset providing audience evaluation metrics from IMDb, including IMDb scores and vote counts. These datasets are combined to enable a comprehensive exploration of both supply-side characteristics, such as content type, release year, and country of production, and demand-side indicators related to audience reception and popularity.

The study adopts an exploratory and descriptive analytical approach. It examines how the size and composition of Netflix's catalogue have changed over time, with a particular focus on the balance between movies and TV shows and the geographical distribution of content production. In addition, the analysis explores patterns in IMDb scores and vote counts to assess how audience evaluation and engagement differ across content types.

Rather than aiming to establish causal relationships, the project seeks to identify trends, distributions, and structural patterns within Netflix's catalogue. By combining descriptive statistics and visual exploration, the analysis provides an overview of Netflix's content evolution and audience response. The findings contribute to a broader understanding of how a global streaming platform curates its content portfolio and how different types of content are perceived by audiences.

## Introduction

Over the past two decades, digital streaming platforms have fundamentally transformed the way audiovisual content is produced, distributed, and consumed. Among these platforms, Netflix has emerged as a dominant global actor, reshaping not only viewing habits but also the structure of the entertainment industry itself. With a presence in more than 190 countries and a continuously expanding catalogue of movies and television shows, Netflix offers a unique case for data-driven analysis of global media production and consumption.

The rapid growth of Netflix has been accompanied by significant strategic shifts. These include a transition from licensed content to original productions, an increased emphasis on episodic television formats, and a strong focus on international markets. As a result, Netflix's content catalogue reflects both technological change and broader cultural and economic trends within the global media landscape. Understanding how this catalogue has evolved over time provides valuable insights into platform strategy, audience targeting, and the globalization of media content.

This project aims to analyse the development of Netflix's content catalogue using publicly available datasets and reproducible data analysis methods implemented in R and Quarto. By combining metadata on Netflix titles with audience evaluation data from IMDb, the analysis seeks to explore patterns of content growth, composition, geographical diversity, and audience reception. Rather than focusing on individual titles, the study adopts a macro-level perspective, treating Netflix's catalogue as a dynamic dataset that evolves over time.

## A brief History of Netflix

Netflix is a subscription-based video-on-demand streaming service founded in the United States in 1997. Initially, the company operated as a mail-based DVD rental service, offering customers access to a broad catalogue of films without the physical limitations of traditional video rental stores. This business model allowed Netflix to compete directly with established retailers such as Blockbuster by leveraging a so-called "long tail" strategy, in which a large number of niche titles collectively generated significant demand.

In the mid-2000s, advances in internet infrastructure, including increased bandwidth and reduced data transfer costs, enabled Netflix to fundamentally transform its business model. The introduction of online streaming marked a decisive shift away from physical media and positioned Netflix at the forefront of digital content delivery. This transition not only reduced distribution costs but also allowed the company to collect detailed, real-time data on user behaviour.

Over time, Netflix increasingly invested in original content production, launching its first original series in the early 2010s. This strategic move reduced dependence on external content providers and enabled greater control over intellectual property. Simultaneously, Netflix expanded aggressively into international markets, commissioning locally produced content in

multiple languages and targeting diverse regional audiences. These developments have resulted in a highly heterogeneous catalogue, making Netflix an ideal subject for empirical analysis of global content strategies.

## **An Analytical View**

A defining characteristic of Netflix's success lies in its systematic use of data to inform strategic decisions. Unlike traditional broadcasters, Netflix operates in a digital environment that enables continuous monitoring of user interactions, such as viewing duration, content completion rates, and engagement patterns. These data allow Netflix to optimise content recommendations, inform commissioning decisions, and evaluate audience response at scale.

From an analytical perspective, Netflix represents a data-rich ecosystem in which content characteristics, production contexts, and audience evaluations intersect. By analysing metadata such as release year, content type, country of production, and runtime alongside IMDb scores and vote counts, it becomes possible to explore both supply-side and demand-side dimensions of the platform. IMDb scores serve as a proxy for perceived content quality, while vote counts provide an indicator of popularity and audience engagement.

This project adopts a quantitative, exploratory approach to analyse these dimensions. Using structured datasets and reproducible workflows, the analysis seeks to identify trends, distributions, and relationships within Netflix's catalogue. The emphasis is not on causal inference, but on descriptive and comparative insights that contribute to a broader understanding of how a global streaming platform curates and positions its content over time. ## Research Questions

How has Netflix's content catalogue evolved over time in terms of content type and country of production?

The main research question aims to explore the structural development of Netflix's content catalogue over time. As one of the largest global streaming platforms, Netflix continuously adapts its content strategy to changing market conditions, audience preferences, and international expansion goals. By analysing trends in content growth, content type, and country of production, this study seeks to provide a comprehensive overview of how Netflix has shaped its catalogue and positioned itself within the global entertainment industry.

To address this overarching research question in a structured manner, the following sub-questions are formulated. Each sub-question focuses on a specific aspect of Netflix's content strategy and contributes to a deeper understanding of the platform's evolution.

## **1. How does the relationship between release year and date added reflect Netflix's strategy in balancing newly produced content with the acquisition and digitisation of older titles?**

This question examines the relationship between the original release year of a title and the date it was added to Netflix in order to better understand the platform's content acquisition strategy. By analysing how recently released productions compare to older titles in terms of when they appear in the catalogue, the study aims to identify whether Netflix primarily prioritises contemporary content or actively invests in digitising and licensing older films and series.

Understanding this relationship provides insight into how Netflix balances freshness and relevance with catalogue depth. A strong concentration of recently released titles may indicate a focus on keeping the platform up to date and competitive, whereas the inclusion of older titles suggests a strategy of enriching the library with classics and historically significant content.

This analysis helps to contextualise subsequent findings by offering a structural perspective on how the platform curates its catalogue over time, not just in terms of quantity, but in terms of the age profile of the content being added.

---

## **2. How has the balance between movies and TV shows evolved on Netflix?**

This question focuses on changes in the composition of Netflix's catalogue, specifically the relative proportions of movies and TV shows over time. Movies and TV shows differ significantly in production costs, audience engagement, and viewing behaviour, making this distinction particularly relevant for understanding Netflix's strategic focus.

By examining how the balance between these two content types has changed, the study aims to assess whether Netflix has increasingly prioritised episodic content, which often encourages longer viewer engagement, or whether movies continue to play a central role in the platform's offering. This analysis contributes to understanding how Netflix aligns its content portfolio with user consumption patterns and long-term retention strategies.

---

### **3. Which countries contribute the most content to Netflix, and how do their contributions differ across content genres?**

This question examines the geographical composition of Netflix's content catalogue with a particular focus on genre distribution. While overall content volume provides an indication of production intensity, analysing genres allows for a more nuanced understanding of how different countries contribute to the diversity of Netflix's offerings.

By investigating which countries dominate the production of specific genres, this question aims to identify potential regional specialisations and production patterns. Such patterns may reflect cultural preferences, industry strengths, or strategic decisions by Netflix to invest in particular types of content in certain markets. This analysis contributes to a deeper understanding of Netflix's global content strategy and its approach to genre diversification across regions.

---

### **4. How are IMDb scores and IMDb vote counts distributed across Netflix titles, and how do they differ between movies and TV shows?**

This question focuses on audience evaluation and popularity of Netflix titles using IMDb data. IMDb scores provide a measure of perceived content quality, while IMDb vote counts reflect the level of audience engagement and visibility of a title. Analysing both variables allows for a more comprehensive understanding of how Netflix content is received by viewers.

By examining the distribution of IMDb scores and vote counts across the catalogue, this question aims to identify general patterns in audience ratings and popularity. Furthermore, comparing movies and TV shows helps to assess whether episodic content and standalone productions differ in terms of viewer reception and engagement levels. This analysis contributes to understanding how content type influences both perceived quality and audience participation.

---

Together, these research questions provide a structured framework for analysing the evolution of Netflix's content catalogue. They allow for a comprehensive examination of growth patterns, content composition, geographical diversity, and audience targeting, thereby supporting a holistic understanding of Netflix's content strategy over time.

## Data Overview

### Data Sample

#### Dataset “Netflix Movies and TV Shows”

```
# A tibble: 8,807 x 3
  title                                country      duration
  <chr>                                <chr>      <chr>
1 Dick Johnson Is Dead                United States 90 min
2 Blood & Water                       South Africa 2 Season~
3 Ganglands                           <NA>         1 Season
4 Jailbirds New Orleans                <NA>         1 Season
5 Kota Factory                        India        2 Season~
6 Midnight Mass                       <NA>         1 Season
7 My Little Pony: A New Generation    <NA>         91 min
8 Sankofa                             United States, Ghana, Burkina Faso~ 125 min
9 The Great British Baking Show       United Kingdom 9 Season~
10 The Starling                       United States 104 min
# i 8,797 more rows
```

#### Dataset “Netflix IMDB Scores”

```
# A tibble: 5,283 x 3
  title                                type  imdb_score
  <chr>                                <chr>    <dbl>
1 Taxi Driver                        MOVIE      8.3
2 Monty Python and the Holy Grail    MOVIE      8.2
3 Life of Brian                      MOVIE      8
4 The Exorcist                       MOVIE      8.1
5 Monty Python's Flying Circus       SHOW      8.8
6 Dirty Harry                       MOVIE      7.7
7 My Fair Lady                      MOVIE      7.8
8 The Blue Lagoon                   MOVIE      5.8
9 Bonnie and Clyde                  MOVIE      7.7
10 The Professionals                 MOVIE      7.3
# i 5,273 more rows
```

## Join data

The selection of an appropriate join strategy was a critical component of the data acquisition phase. To ensure the integrity and completeness of the final dataset, the team opted for a Full Outer Join.

This approach was chosen for the following reasons:

- Data Preservation: It ensures that no information is inadvertently lost during the merging process, even if the data is incomplete in one of the source tables.
- Holistic Analysis: It allows for a comprehensive view of the data, enabling the identification of gaps or discrepancies between the merged datasets.

```
# A tibble: 10,343 x 5
```

title	country	duration	type.y	imdb_score
<chr>	<chr>	<chr>	<chr>	<dbl>
1 Dick Johnson Is Dead	United States	90 min	MOVIE	7.4
2 Blood & Water	South Africa	2 Seaso~	<NA>	NA
3 Ganglands	<NA>	1 Season	SHOW	7
4 Jailbirds New Orleans	<NA>	1 Season	SHOW	6.6
5 Kota Factory	India	2 Seaso~	SHOW	9.3
6 Midnight Mass	<NA>	1 Season	SHOW	7.7
7 My Little Pony: A New Generation	<NA>	91 min	MOVIE	6.8
8 Sankofa	United States, G~	125 min	MOVIE	7
9 The Great British Baking Show	United Kingdom	9 Seaso~	SHOW	8.6
10 The Starling	United States	104 min	MOVIE	6.3

```
# i 10,333 more rows
```

Due redudant columns (e.g. description.x, desxription.y) , we applied a cleaning step.

## Exploratory Data Analysis

### Statistics

#### Numerical Values

release_year.x	runtime	imdb_score	imdb_votes
Min. :1925	Min. : 0.0	Min. :1.500	Min. : 5
1st Qu.:2013	1st Qu.: 45.0	1st Qu.:5.800	1st Qu.: 521
Median :2017	Median : 87.0	Median :6.600	Median : 2279
Mean :2014	Mean : 79.2	Mean :6.533	Mean : 23407
3rd Qu.:2019	3rd Qu.:106.0	3rd Qu.:7.400	3rd Qu.: 10144

Max.	:2021	Max.	:235.0	Max.	:9.600	Max.	:2268288
NA's	:1493	NA's	:5060	NA's	:5060	NA's	:5076

## Categorical Values

\$type.x

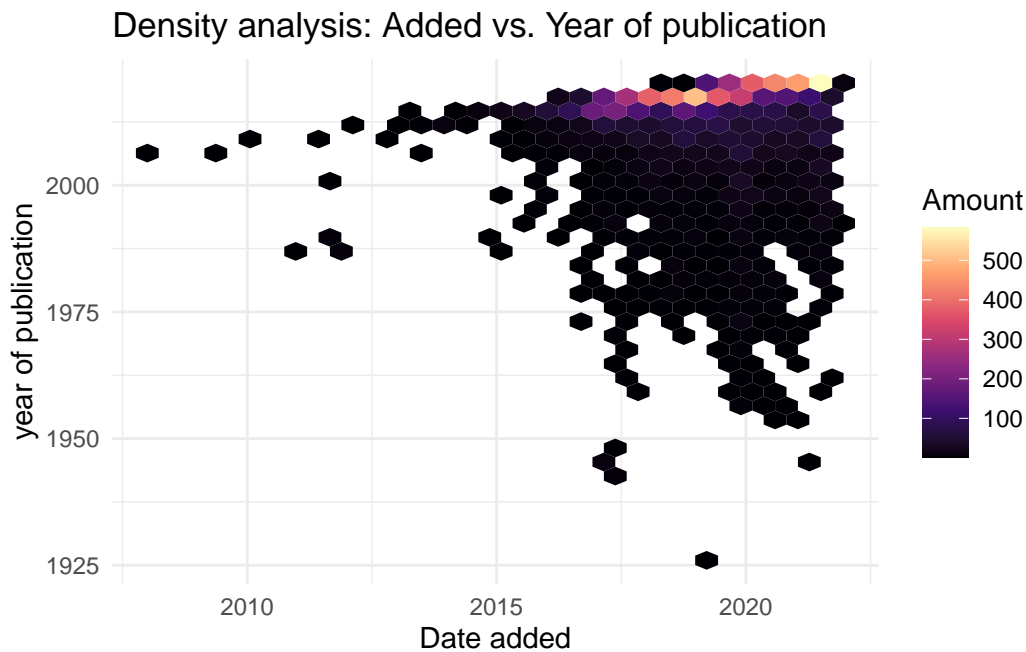
Movie	TV	Show
6158	2692	

\$age\_certification

G	NC-17	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y	TV-Y7
105	13	238	424	548	436	72	792	172	94	104

## Statistical Analysis & Methodology

### Release Date vs. Date added



The density plot illustrates the relationship between the year a title was released and the date it was added to Netflix. A strong concentration is visible in the upper-right area of the plot, indicating that many titles added to the platform were released only a few years earlier. This



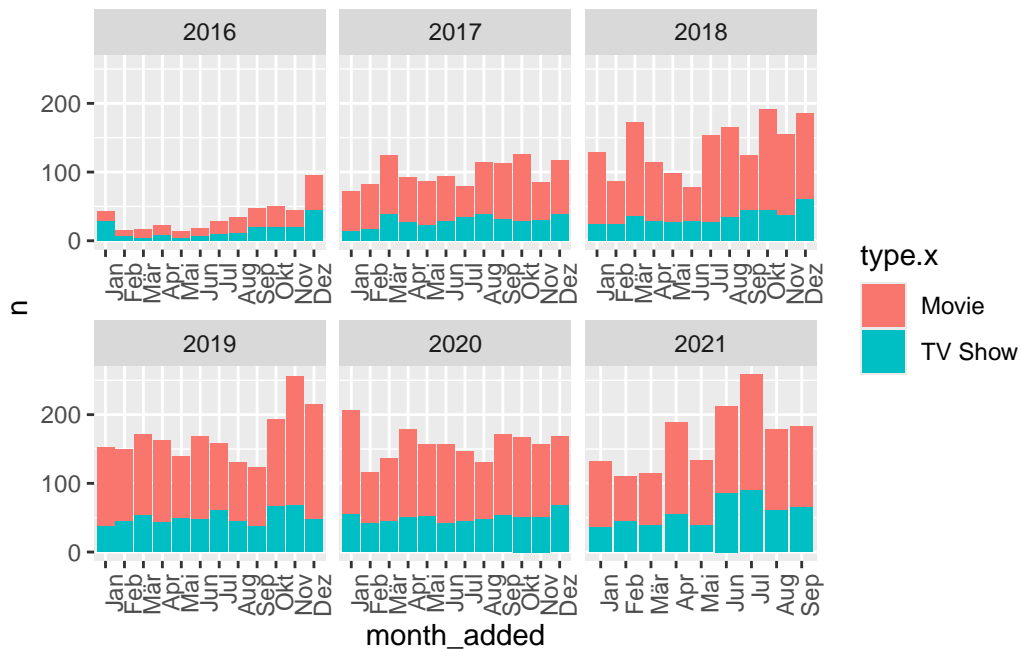
pattern suggests that Netflix places a strong emphasis on acquiring and distributing relatively recent productions.

At the same time, the presence of titles with much older release years shows that the platform also incorporates older films and series into its catalogue. However, these titles appear less frequently and are more sparsely distributed, indicating that digitisation or acquisition of classic content plays a secondary role compared to the continuous addition of newer material.

Overall, the visualisation suggests that Netflix's content strategy prioritises contemporary titles while selectively enriching its catalogue with older productions, thereby maintaining a balance between current relevance and historical depth.

## Balance plot

To examine how the relative contribution of movies and TV shows has changed over time, the monthly number of titles added to the platform is visualised separately for each year.



The balance plot illustrates how the number of movies and TV shows added to Netflix has evolved over time, broken down by month and year. Across all years, movies consistently account for a larger share of newly added titles, indicating that films remain a core component of the platform's catalogue.

At the same time, the presence of TV shows increases steadily over the observed period. In several months, especially in later years, the number of added series becomes more substantial, suggesting a growing emphasis on episodic content. This pattern aligns with Netflix's strategic

focus on long-form storytelling formats that can encourage sustained viewer engagement and retention.

Overall, the visualisation indicates that while movies continue to dominate in terms of volume, TV shows represent an increasingly important component of the platform's content mix. The evolving balance between these two content types reflects a gradual shift toward a more series-driven catalogue without fully replacing the importance of film-based offerings.

## Country and type distribution

In the joined dataset, several variables contain multiple values within a single cell. In particular, the variables *country* and *listed\_in* (genre) often include more than one entry, separated by commas. This structure reflects the fact that a single Netflix title can be associated with multiple producing countries and multiple genres. However, such a format is not suitable for aggregation and frequency analysis in its raw form.

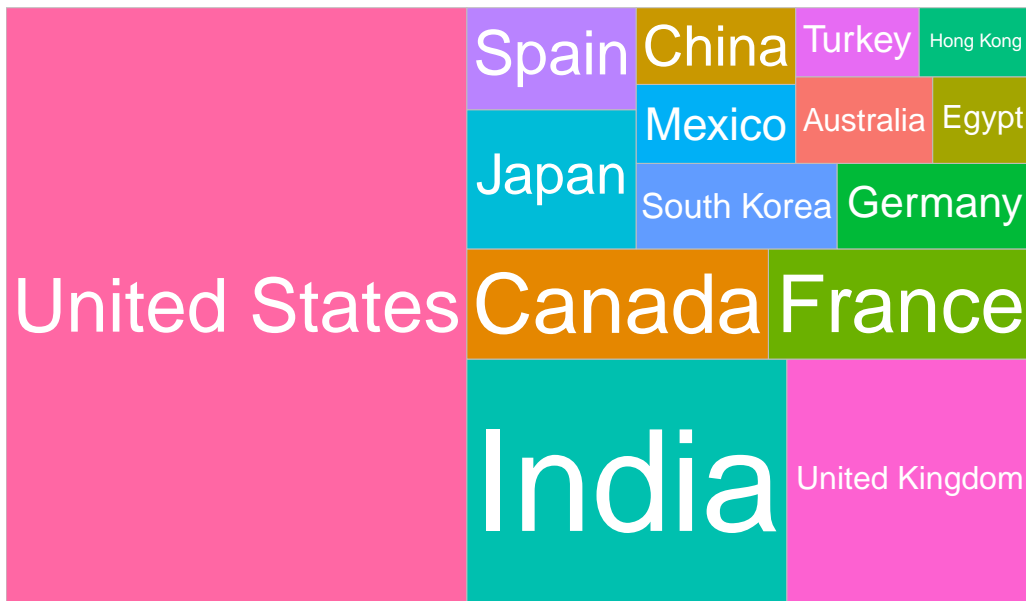
To address this issue, the values in these columns are first separated into individual entries. Each country or genre is extracted and treated as a separate observation. This transformation allows for a more accurate representation of how frequently each country and each genre appears across the Netflix catalogue. After splitting the values, records with missing or empty entries are removed to ensure data quality.

Following this preprocessing step, the data is aggregated by counting the number of occurrences of each country and each genre. To improve readability and focus on the most relevant categories, the analysis is restricted to the top fifteen countries and the top fifteen genres based on their frequency. This approach highlights the dominant contributors while avoiding visual clutter caused by less frequent categories.

The resulting summaries are visualised using treemap plots. In these visualisations, the size of each rectangle represents the relative frequency of a country or genre within the dataset. Treemaps are particularly well suited for this purpose, as they provide an intuitive overview of proportional contributions and allow for easy comparison between categories. These visualisations support an exploratory understanding of the geographical and thematic composition of Netflix's content catalogue.

To visualise the most frequent producing countries and content genres, treemap plots are created. For both variables, the multi-value entries are first expanded so that each country or genre appears as a separate observation. The data is then aggregated by counting the frequency of each category, and the analysis is restricted to the top 15 entries to ensure readability of the visualisations.

Countries (Treemap)



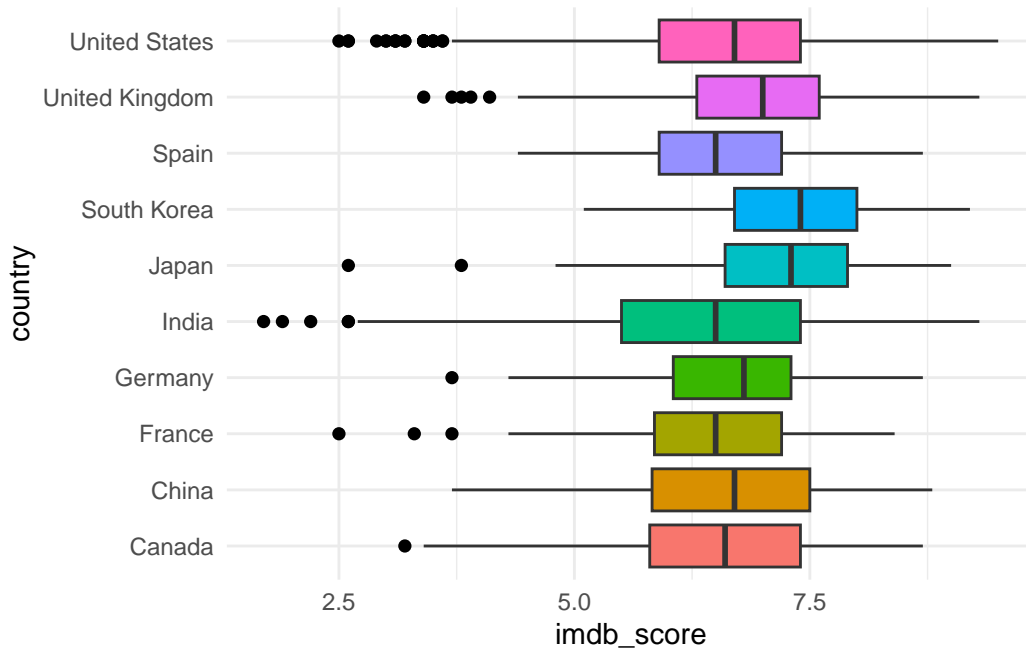
The country treemap suggests that Netflix content production is concentrated in a small number of countries. A few countries occupy the largest areas in the visualisation, indicating that they contribute a substantial share of the titles in the dataset. At the same time, the presence of many smaller rectangles highlights the involvement of a wide range of additional countries, each contributing a smaller number of titles. This pattern reflects a combination of dominant production markets and broader international diversity within Netflix's catalogue.

## Genres (Treemap)



The genre treemap indicates that a limited number of genres account for a large proportion of Netflix's content catalogue. These dominant genres form the core of the platform's offering, while numerous other genres appear less frequently. The distribution suggests that Netflix combines widely popular genres with a variety of more specialised categories, contributing to thematic diversity across the catalogue.

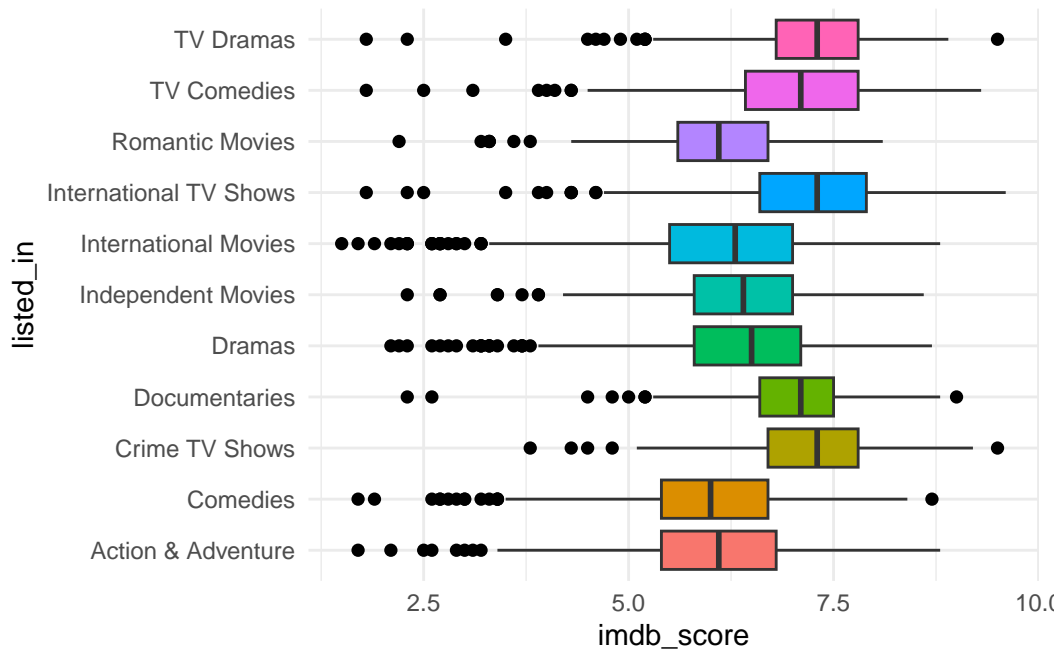
## IMDB Ranking over Countries and Genres



The boxplot illustrates the distribution of IMDB scores across the ten countries that contribute the largest number of titles to the dataset. Rather than focusing on single high-scoring productions, the visualisation highlights the overall quality distribution within each country's output.

The results show that most countries have a relatively similar central tendency, with median scores generally concentrated in the mid-range around 6 to 7. This suggests that perceived content quality does not vary drastically between the major production countries represented on Netflix. At the same time, the spread of the boxplots indicates that each country produces both highly rated and more moderately rated titles.

Some countries display wider variability, which may reflect a broader diversity of genres and production styles. Overall, the figure suggests that Netflix's major content-producing countries contribute comparably in terms of average audience reception, with differences appearing more in variability than in central quality levels.



This visualisation compares the distribution of IMDb scores across the ten most frequent content genres on Netflix. By examining the median, spread, and outliers of each genre, the plot provides insight into how audience evaluations differ depending on thematic category.

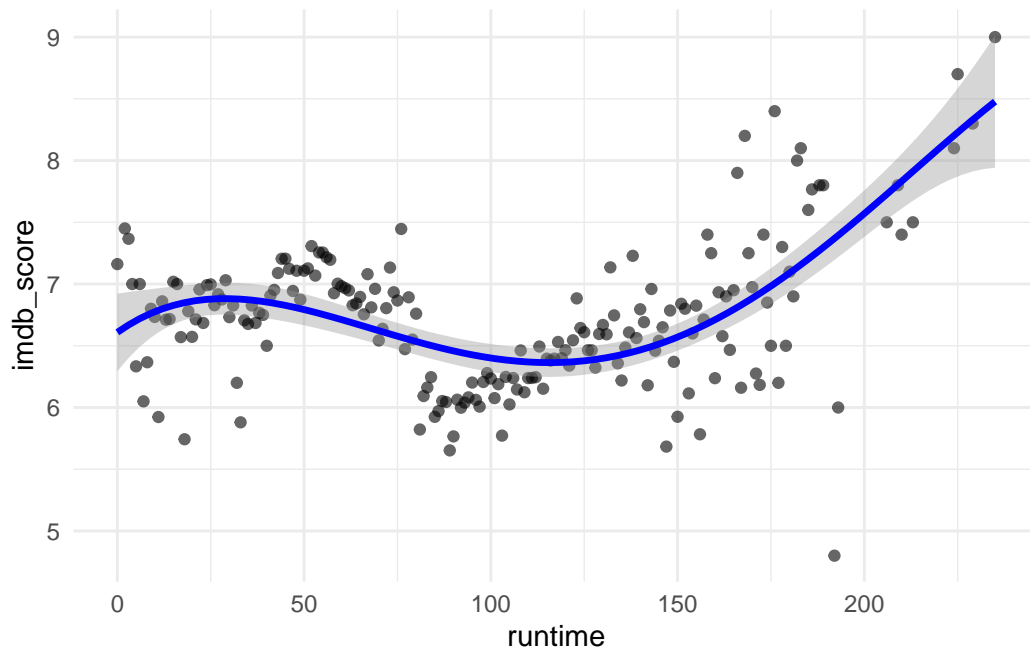
The results indicate that certain genres tend to achieve consistently higher audience ratings, while others show greater variation. In particular, documentary and drama-related categories often exhibit higher median scores, suggesting stronger audience appreciation and perceived quality. In contrast, more entertainment-focused genres such as comedies and action-oriented content show a broader distribution of ratings, reflecting a mix of both highly successful and more average productions.

Overall, the distribution suggests that genre plays an important role in shaping audience perception. While no single genre dominates completely, some categories appear more consistently associated with higher evaluation scores, indicating differences in how viewers respond to different types of storytelling.

## Chapter of Choice: Forecasting Content Quality Based on Runtime

This chapter presents an additional exploratory analysis that was not part of the core course content. While the main sections of the report focus on descriptive insights into Netflix's catalogue, this section introduces a simple predictive modelling approach.

The goal of this analysis is to explore whether structural characteristics of a title — in this case runtime — are associated with differences in audience evaluation, measured by IMDb score. Although this does not represent a formal forecasting model, it provides an example of how regression techniques can be used to identify potential trends and relationships within the dataset.



The forecasting model explores the relationship between runtime and average IMDb score by fitting a smoothed polynomial trend line to the observed data. The scatterplot shows individual titles, while the fitted curve highlights the general pattern across different runtime lengths.

The results suggest a non-linear relationship between runtime and perceived quality. Titles with very short or moderate runtimes tend to cluster around mid-range scores, while longer productions show a gradual increase in average ratings. This pattern may indicate that longer formats, which often allow for more complex storytelling and character development, are associated with slightly higher audience appreciation.

At the same time, the wide dispersion of points demonstrates that runtime alone is not a strong predictor of quality. High- and low-rated titles exist across nearly all runtime ranges. The model therefore provides an exploratory indication of a potential trend rather than a causal relationship, highlighting how runtime may be linked to audience perception without determining it directly. ## Discussion

**Conclusion**

**References**

**Appendix**