# Netflix Content Trends:Exploratory Data Analysis

Prepared by: Soubhanik Patra

Date: July 2025

# 1. Introduction

The global shift in entertainment consumption has made Netflix a major player in the streaming industry. This project performs a structured Exploratory Data Analysis (EDA) of Netflix's Movies and TV Shows dataset. By combining data visualization, feature engineering, and hypothesis testing, we uncover meaningful trends in content type, genre, duration, and ratings over time.

---

# 2. Dataset Overview

- **Source**: Kaggle – Netflix Movies and TV Shows
- **Records**: 8,807 titles
- **Columns**: 12 original features
- **Time Range**: Content from 1925 to 2021

### Key Variables:

| Column | Description |
|---|---|
| `show_id` | Unique identifier |
| `type` | Movie or TV Show |
| `title` | Name of the title |
| `director` | Director's name |
| `cast` | Lead actors/actresses |
| `country` | Country of production |
| `date_added` | Date when it was added to Netflix |
| `release_year` | Year of release |
| `rating` | Maturity rating (e.g., TV-MA, PG-13) |
| `duration` | Duration in minutes or seasons |
| `listed_in` | Genre(s) |
| `description` | Short summary |

---

# 3. Data Cleaning and Feature Engineering

To prepare the data for analysis:

- **Handled missing values**:
  - Replaced nulls in `director`, `cast`, and `country` with `'Unknown'`
  - Dropped rows with null `date_added`
- **Created derived features**:
  - `year_added`: Extracted from `date_added`
  - `duration_mins`: Extracted numeric part of `duration`

- `main_genre`: First genre from `listed_in`
- `primary_country`: First country listed in `country`

```
df['year_added'] = pd.to_datetime(df['date_added']).dt.year
df['duration_mins'] = df['duration'].str.extract('(\d+)').astype(float)
df['main_genre'] = df['listed_in'].str.split(',').str[0]
df['primary_country'] = df['country'].str.split(',').str[0]
```

---

# 4. Exploratory Data Analysis (EDA)

## 🎥 Type Distribution

- Netflix has more **Movies** than TV Shows, but **TV Shows have surged** after 2016.

## 📅 Content Over Time

- A clear spike in content additions is visible between **2016 and 2020**.

## 🌍 Top Countries

- The **United States** leads in content count, followed by **India**.

## 🎭 Genre Distribution

- **Drama**, **Comedy**, and **Documentary** are the most common genres.

## ⏱️ Duration

- Most movies are between **80–100 minutes** long.
- TV Shows typically report their duration as number of **seasons**.

## 🔞 Ratings

- Majority of content is rated **TV-MA** and **PG-13**, targeting mature audiences.

---

# 5. Key Insights

- The number of **TV Shows** has significantly increased post-2016.
- **Drama** and **Comedy** dominate across both movies and shows.
- **India** is the second largest contributor to Netflix’s catalog.
- There’s a noticeable **decline in average movie duration** after 2018.
- Most recent content targets **mature audiences**.

---

## 6. Hypothesis Testing

### 📊 Hypothesis 1:

**"Rating distribution differs between Movies and TV Shows."**

- **Test Used**: Chi-square test of independence
- **Result**: p-value < 0.001
- ✅ **Conclusion**: Statistically significant difference

### 📊 Hypothesis 2:

**"Average movie duration has decreased after 2018."**

- **Test Used**: Independent t-test
- **Result**: p-value < 0.05
- ✅ **Conclusion**: Significant decline in duration after 2018

---

## 7. Conclusion

Netflix’s content strategy has evolved to focus more on **TV Shows**, especially after 2016. The platform’s catalog is shaped by the US and India, leans toward **mature-rated content**, and increasingly favors **shorter movies**. These insights could inform content recommendations, production strategies, or predictive modeling tasks.

---

## 8. Future Work

- Train an ML classifier to predict content type (Movie or TV Show)
- Use clustering to identify genre-based user segments
- Extend to personalized recommendations if viewer data is available
- Perform NLP on descriptions for thematic analysis

---

## 9. Acknowledgments

- Dataset sourced from Shivam Bansal on Kaggle