# Project Documentation: Netflix Content Analysis

## Data visualization

Name: Osama Ahmed Mohamed Ahmed Amer

Id: 20201701525

Department: Bioinformatics – 4

# 1. Introduction

- Purpose of the Project:
  this project explores the Netflix dataset to visualize key trends in its content. The analysis focuses on content types, release patterns, movie durations, ratings, title themes, and leading content-producing countries.

- Dataset Used:
  The "Netflix Movies and TV Shows" dataset (2021 version, from Kaggle) was used. Key attributes analyzed include type, title, country, release_year, rating, and duration.

- Column in the data
  **show_id:**
  A unique identifier for each movie or TV show listing.
  **type:**
  Categorizes the content as either a "Movie" or a "TV Show".
  **title:**
  The name of the movie or TV show.
  **director:**
  The name(s) of the person(s) who directed the content.
  **cast:**
  A list of the main actors and actresses in the content.
  country:

The country or countries where the content was produced.

**date_added:**

The date when the content was added to the Netflix library.

**release_year:**

The year the content was originally released (e.g., in theaters, on its original network).

**rating:**

The age-based or content-based rating (e.g., TV-MA, PG-13, R).

**duration:**

The length of the content. For movies, it's usually in minutes (e.g., "121 min"). For TV shows, it's often the number of seasons (e.g., "2 Seasons").

**listed_in:**

The genre(s) the content is categorized under (e.g., "Dramas, International Movies", "Comedies").

**description:**

A brief summary or synopsis of the content.

```
Column types:
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
description             object
month_added           category

Missing values per column:
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added         0
release_year       0
rating             4
duration           3
listed_in          0
description        0
month_added        0
```

- Tools Used:

  Analysis and visualization performed in R
  tidyverse (for ggplot2 and data manipulation), lubridate (dates), tidytext, and wordcloud (text analysis).

# 2. Data Preprocessing

- Missing Value Handel:
  Categorical missing values (e.g., in country, rating) were labeled "Missing." Rows with missing duration were removed.

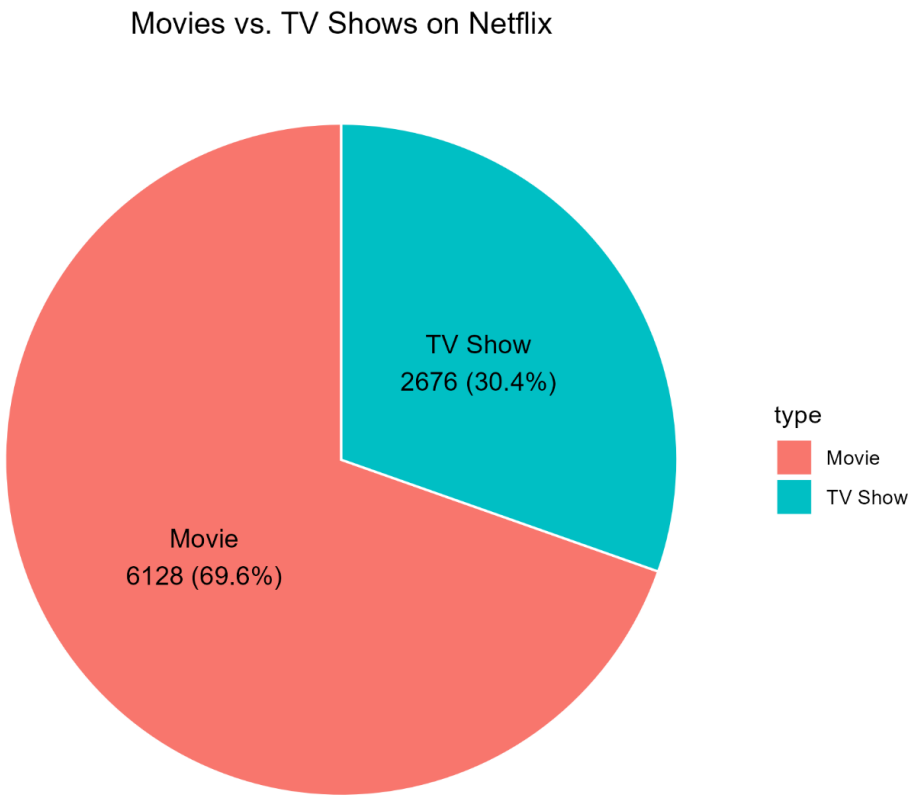
- Duration Standardization:
  The duration column was parsed into numerical values (duration_int) and standardized units (duration_unit like "min" or "Season").

- Date Processing:
  date_added was converted to a date format. Missing date_added entries were imputed using release_year (assuming Jan 1st). month_added and year_added were extracted.

- Text Preparation:
  the title word cloud, common stop words were removed to highlight the important terms.

# 3. visualizations Plots

## 1. Content Type Pie Chart (Movies vs. TV Shows)



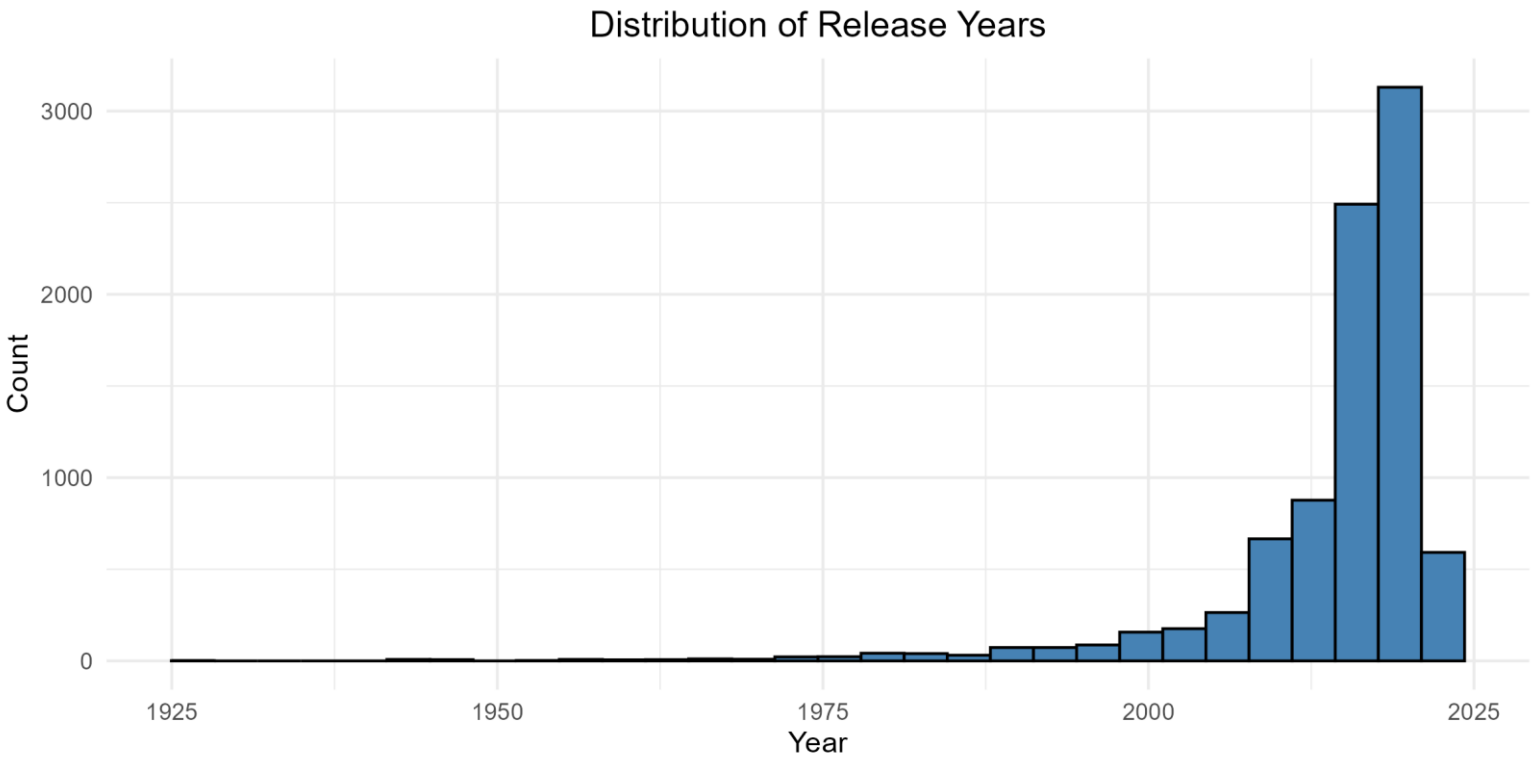

- Why this chart?

  A pie chart clearly shows parts of a whole, ideal for comparing just 'Movies' and 'TV Shows.

- Observations

  Netflix is mostly movies (almost 70%), with TV shows making up the rest. Movies dominate the library.

## 2. Release Year Histogram (Distribution of Release Years)



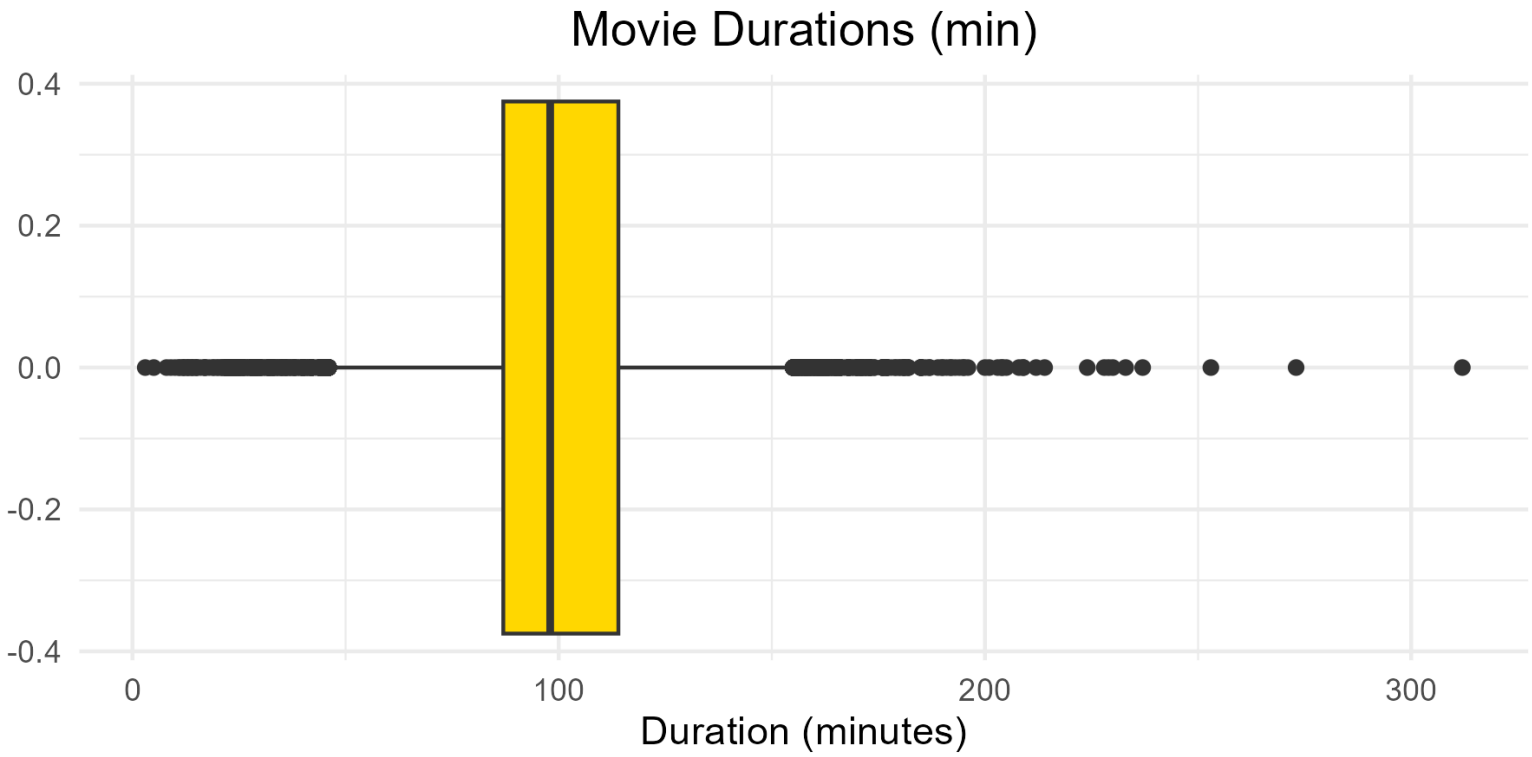

- Why this chart?

  A histogram shows content frequency by release year, revealing production trends over time.

- Observations:

  Most content is recent, spiking after 2010 and peaking around 2018-2020. Netflix favors newer releases.

# 3.Movie Duration Boxplot



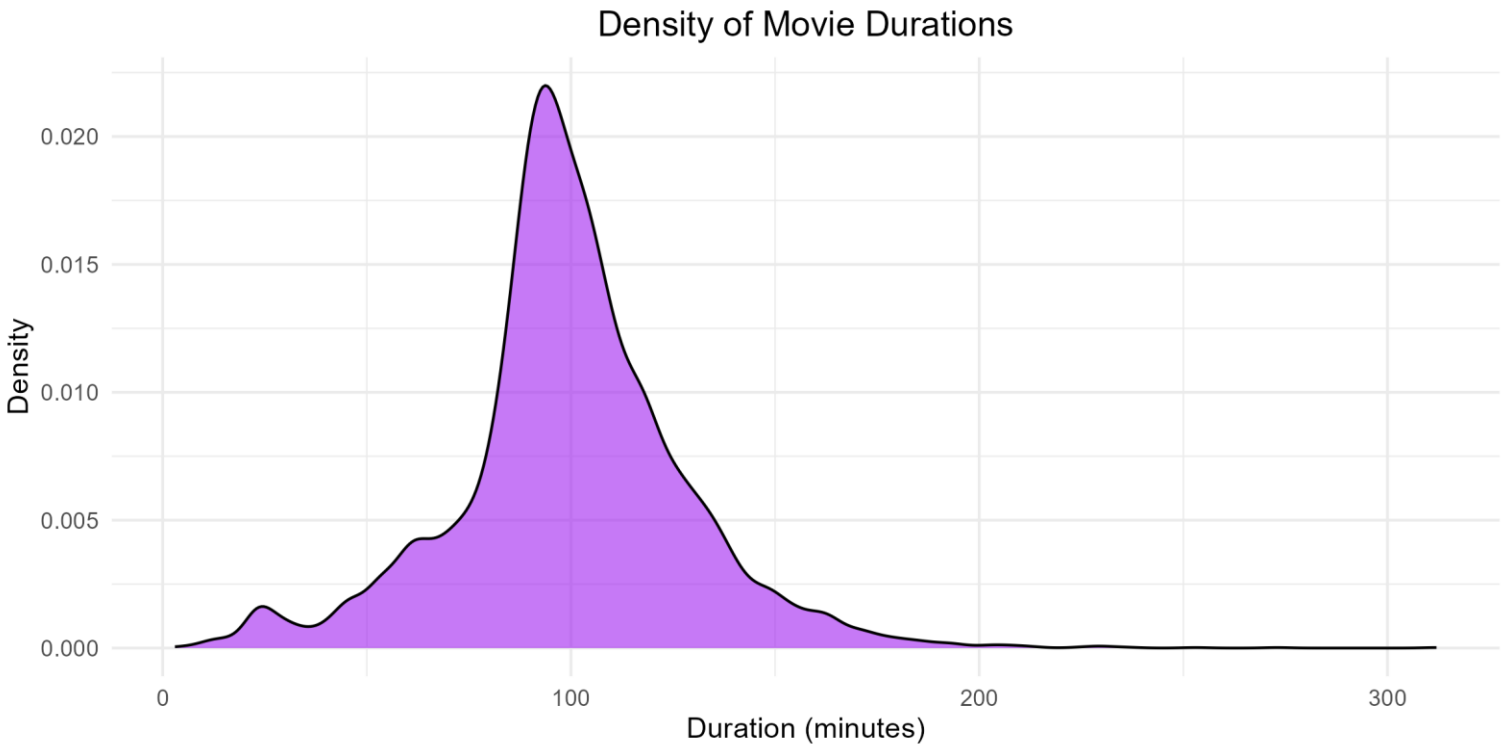

- Why this chart?

  A boxplot quickly shows typical movie lengths, the middle range, and any unusually long films.

- Observations:

  Most movies are 90-120 minutes. The average is just under 100 minutes, with outliers for longer films.

# 4. Density Plot of Movie Durations



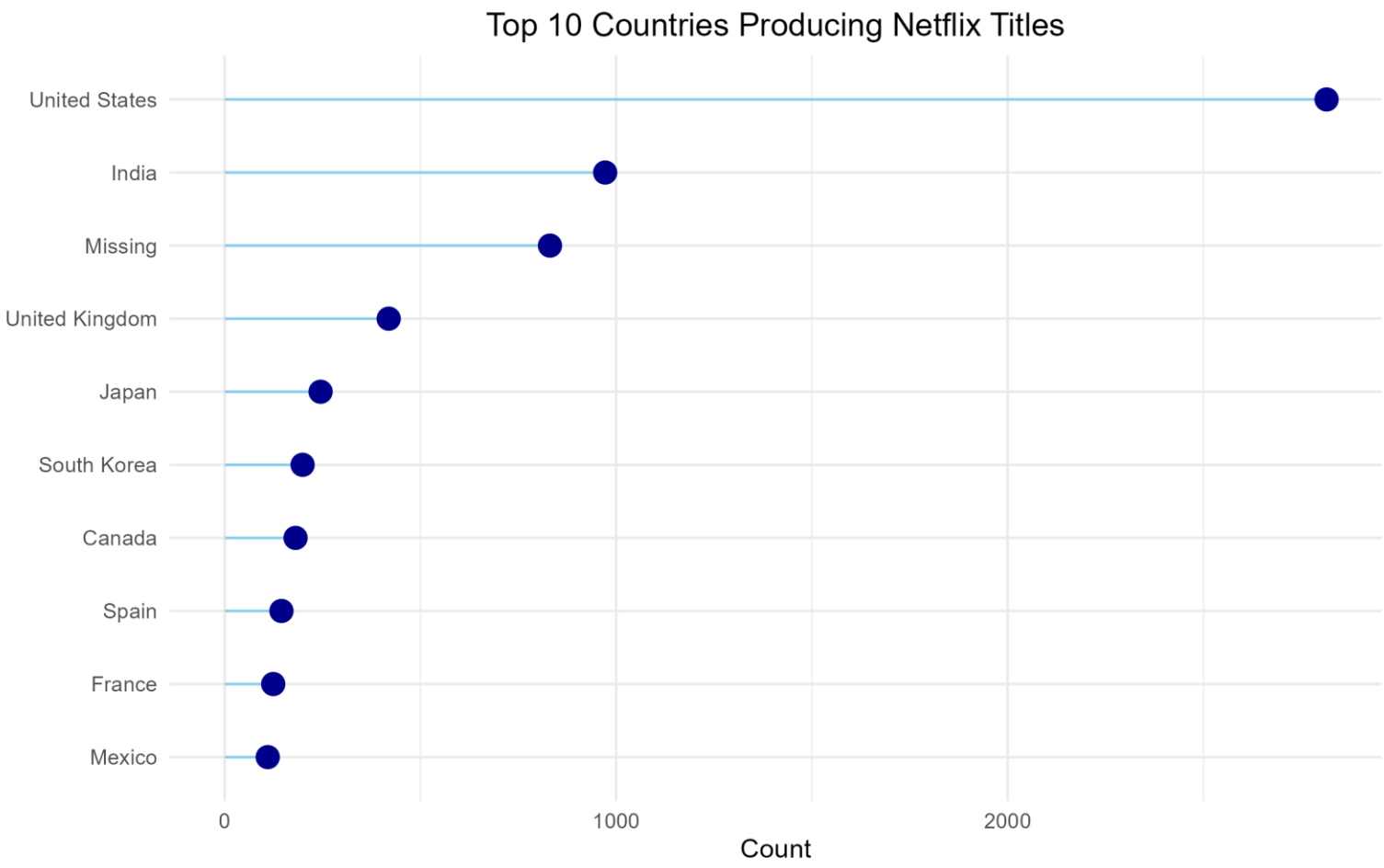

- Why this chart?

  This density plot gives a smooth view of movie duration spread, highlighting common lengths.

- Observations:

  Confirms a peak around 90-100 minutes, with fewer movies as they get longer, showing a right skew.

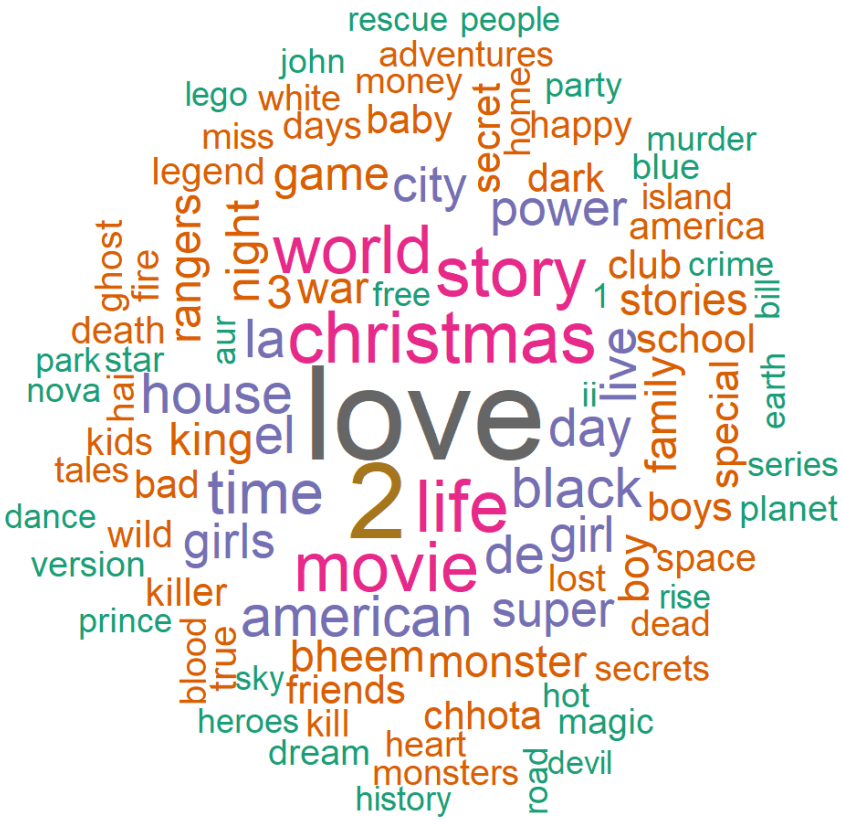
## 5. Top 10 Countries Lollipop Chart (Producing Netflix Titles)




- Why this chart?
  A lollipop chart cleanly ranks top countries, making content origin comparisons easy.

- Observations:
  The US produces most content, followed by India. 'Missing' country data is surprisingly common.

# 6. Word Cloud (From Titles)



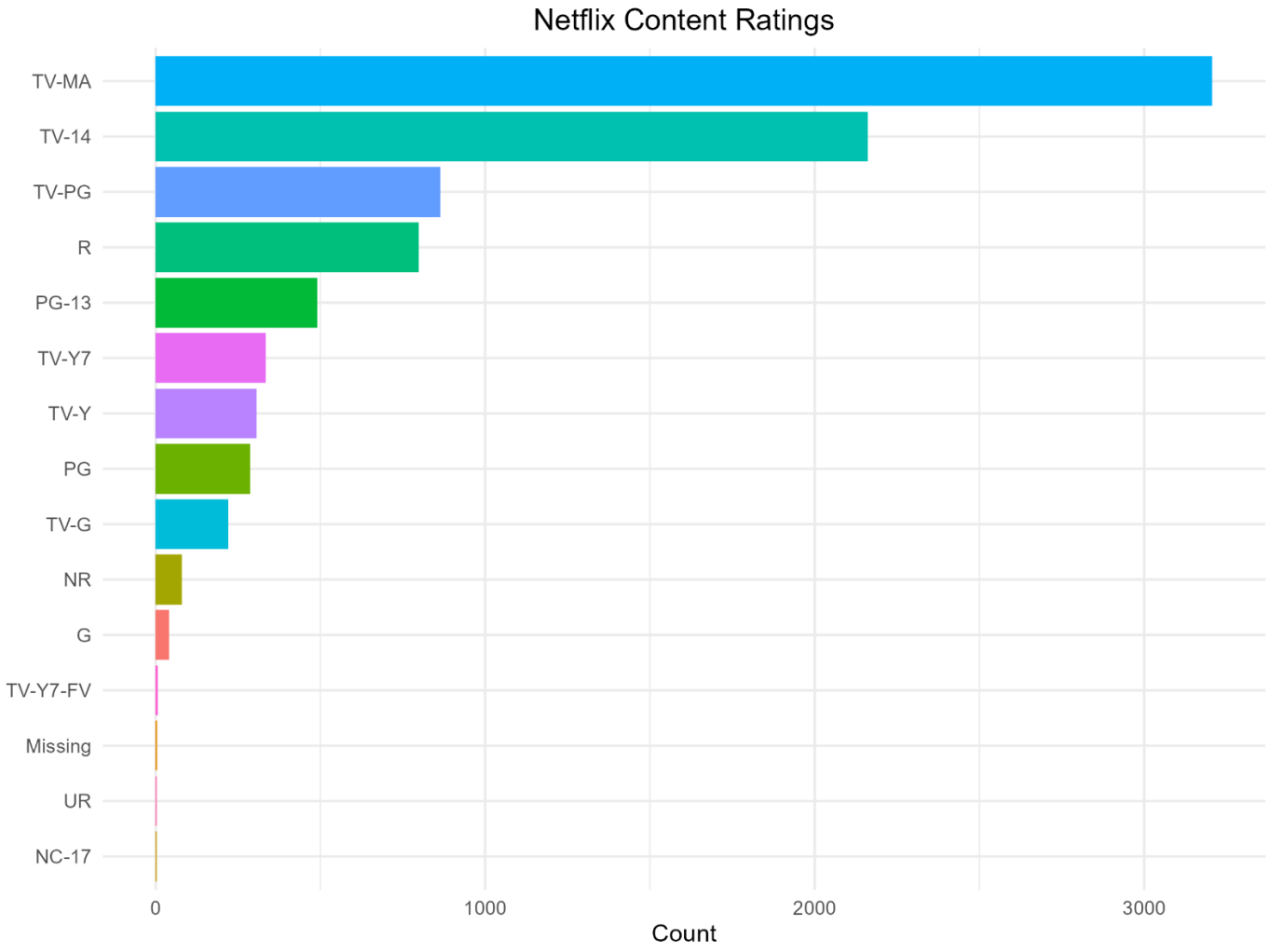

- Why this chart?

  A word cloud visually shows frequent title words, giving a quick feel for common themes.

- Observations:

  'Love' and 'Christmas' are prominent, suggesting romance and holiday themes are popular. 'World,' 'Story,' and '2' also stand out.

## 8. Rating Distribution Bar Chart (Netflix Content Ratings)




- Why this chart?

  A bar chart effectively compares counts across distinct rating categories, showing which are most common.

- Observations:

  TV-MA' is the most common rating, then 'TV-14' and 'TV-PG.' Netflix heavily caters to mature audiences.

## 4. Conclusion: What We Learned About Netflix

- Lots of Movies, Mostly New Stuff
  Netflix has way more movies than TV shows. They also really like to keep things fresh, with most of their content made in the last 10-15 years.

- Standard Movie Night Length
  Most movies run about an hour and a half to two hours, which is pretty typical.

- Made in the USA (and India):
  The US makes the most content, but India is a big player too. Interestingly, a lot of titles are missing country info.

- Love, Christmas, and Adventure
  Common words in titles like "Love," "Christmas," "World," and "Story" suggest these themes are really popular on the platform.

- Mostly for Grown-Ups and Older Teens:
  The "TV-MA" rating (for mature audiences) is the most common, meaning a big chunk of Netflix is aimed at adults.