Netflix Genre Analysis

Mustafa Taha Disbudak

2024-12-11

Table of contents

0.1 Introduction

This report provides an analysis of the distribution of well-known genres within Netflix's catalog, the top cast appearances, and trends in Turkish content over time. The data source is the TidyTuesday project, which is a weekly data activity organized by the R for Data Science (R4DS) community. The dataset originally comes from the Flixable service and was made publicly available through the TidyTuesday Github repository.

```
# Define our genres of interest which are the most well know genres
genres_of_interest <- c(
  "crime", "thriller",
  "comedy",
  "drama",
  "action", "adventure",
  "romantic", "romance",
  "science fiction", "sci-fi", "fantasy",
  "documentary",
  "horror"
)

# Function to clean and split genres
clean_genres <- function(x) {
  x |>
    str_split(",") |>
    unlist() |>
    str_trim() |>
    str_remove_all("(?i)movies|tv shows") |>
    str_trim() |>
    str_to_lower()
```

```
}

# Replace case_when with if/else logic for assigning genre_category
assign_genre_category <- function(g) {
  if (str_detect(g, "(?i)crime|thriller")) {
    "Crime/Thriller"
  } else if (str_detect(g, "(?i)comedy")) {
    "Comedy"
  } else if (str_detect(g, "(?i)drama")) {
    "Drama"
  } else if (str_detect(g, "(?i)action|adventure")) {
    "Action & Adventure"
  } else if (str_detect(g, "(?i)romantic|romance")) {
    "Romantic Comedies"
  } else if (str_detect(g, "(?i)science fiction|sci-fi|fantasy")) {
    "Science Fiction & Fantasy"
  } else if (str_detect(g, "(?i)documentary")) {
    "Documentary"
  } else if (str_detect(g, "(?i)horror")) {
    "Horror"
  } else {
    "Other"
  }
}

# Analyze genres using the new function instead of case_when
genre_counts <- netflix |>
  mutate(genres = map(listed_in, clean_genres)) |>
  unnest(genres) |>
  mutate(genre_category = map_chr(genres, assign_genre_category)) |>
  filter(genre_category != "Other") |>
  distinct(show_id, genre_category) |>  # Count each title only once per category
  count(genre_category, sort = TRUE)

# Print results
print(genre_counts)
```

```
# A tibble: 7 x 2
  genre_category               n
  <chr>                    <int>
1 Drama                     2810
2 Crime/Thriller             958
3 Action & Adventure         871
4 Romantic Comedies          864
5 Comedy                     381
```

```
6 Horror                        381
7 Science Fiction & Fantasy     294
```

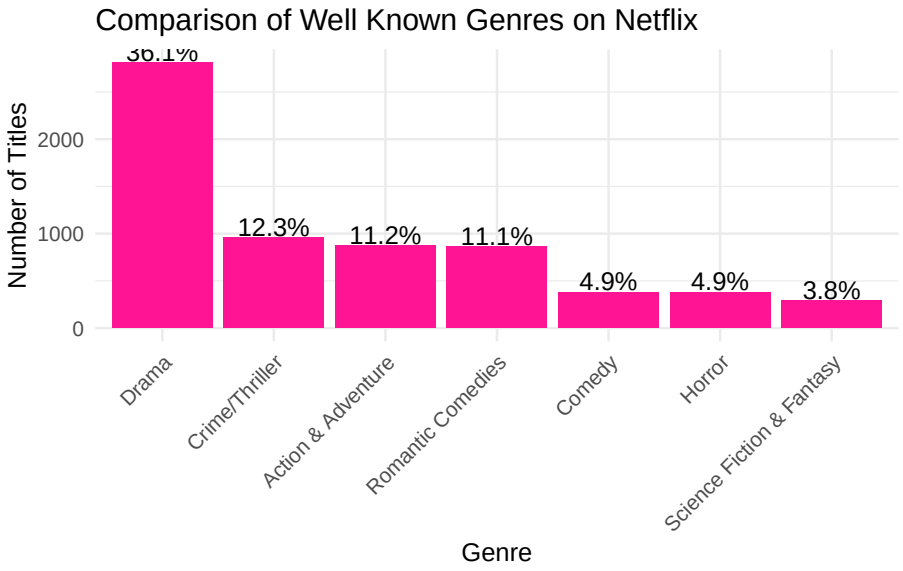
```r
# Calculate total titles
total_titles <- n_distinct(netflix$show_id)
print(paste("Total number of unique titles:", total_titles))
```

```
[1] "Total number of unique titles: 7787"
```

```r
# Calculate percentage for each genre
genre_counts <- genre_counts |>
  mutate(percentage = n / total_titles * 100)

# Print results with percentages
print(genre_counts)
```

```
# A tibble: 7 x 3
  genre_category                n percentage
  <chr>                     <int>      <dbl>
1 Drama                      2810      36.1
2 Crime/Thriller              958      12.3
3 Action & Adventure          871      11.2
4 Romantic Comedies           864      11.1
5 Comedy                      381       4.89
6 Horror                      381       4.89
7 Science Fiction & Fantasy   294       3.78
```

```r
# Create a bar plot for the genre distribution with rotated x-axis labels
ggplot(genre_counts, aes(x = reorder(genre_category, -n), y = n)) +
  geom_bar(stat = "identity", fill = "deeppink") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.1) +
  labs(
    title = "Comparison of Well Known Genres on Netflix",
    x = "Genre",
    y = "Number of Titles"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate labels
```



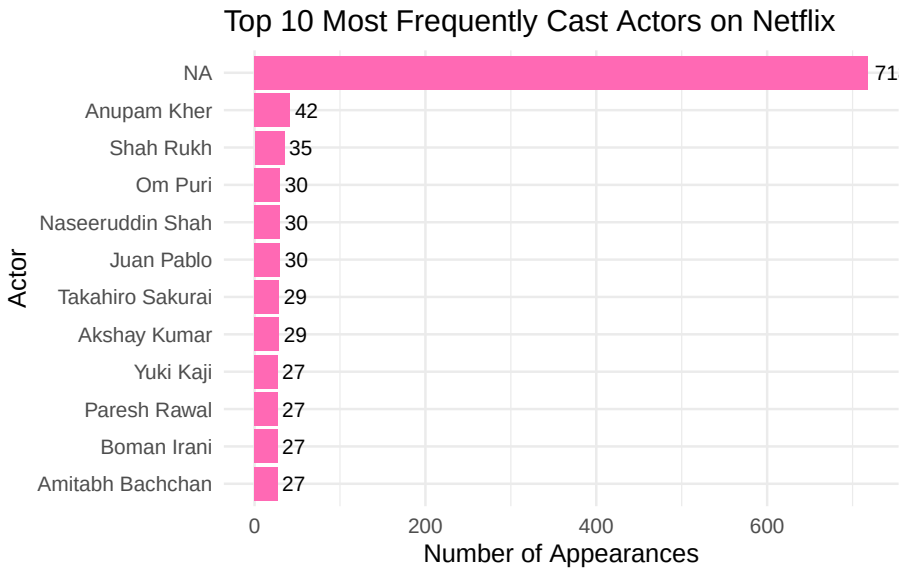
The graph illustrates the distribution of genres within Netflix's library, highlighting a significant dominance of the drama genre, which accounts for over 36% of titles. This prevalence is understandable, as dramas often provide the foundation for captivating stories and frequently overlap with other genres, such as science fiction, romance, or thrillers. The substantial share of dramas suggests that Netflix prioritizes these versatile and emotionally engaging stories to attract a wide audience.

In addition to dramas, the Crime/Thriller and Action & Adventure genres also hold notable shares, each comprising around 12% of the content. These genres likely appeal to viewers seeking suspense and high-stakes plots, drawing consistent interest across various demographics. Furthermore, Romantic Comedies represent 11.1% of the library, catering to audiences looking for light-hearted, feel-good content.

Horror genre hovers around only 5%. This is not surprising to me, as an horror fan I always have an hard time finding a good horror show on Netflix due to comparetavily limited options.

```
# Extract and count actors
actor_counts <- netflix %>%
  mutate(actors = str_extract_all(cast, "\\b[A-Z][a-z]+ [A-Z][a-z]+\\b")) %>%
  unnest(actors) %>%
  count(actors, name = "n", sort = TRUE) %>%  # Corrected the count function
  slice_max(n, n = 10)  # Changed slice_top_n to slice_max

# Plot top 10 actors
ggplot(actor_counts, aes(x = reorder(actors, n), y = n)) +
```

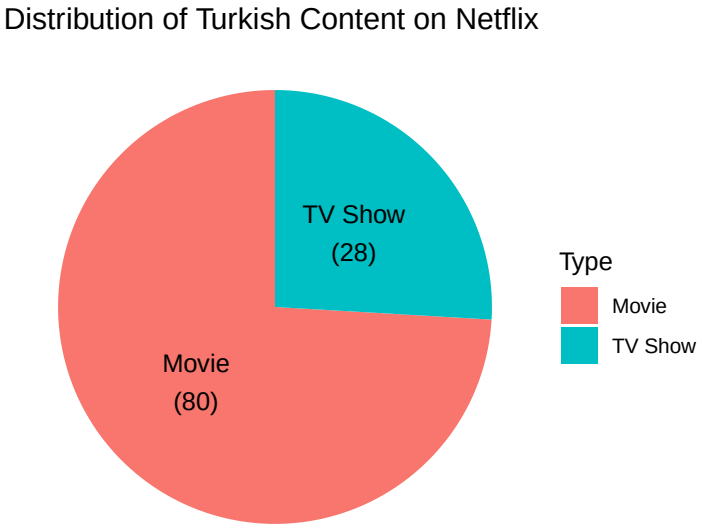
```
  geom_col(fill = "hotpink") +
  coord_flip() +
  labs(title = "Top 10 Most Frequently Cast Actors on Netflix",
       x = "Actor",
       y = "Number of Appearances") +
  theme_minimal() +
  geom_text(aes(label = n), hjust = -0.2, size = 3)
```



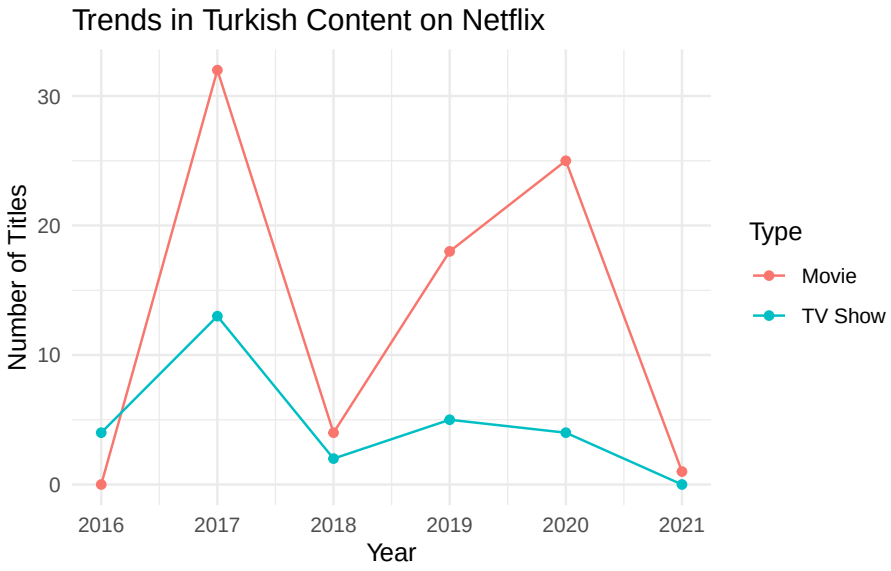
According to the graph the cast information was missing for 718 titles on Netflix which indicated the data we have is not of the highest quality. Anupam Kher is the most frequently cast actor on Netflix with 42 apperances.

```
# Extract shows from Turkey
turkish_shows <- netflix %>%
  filter(str_detect(country, "Turkey")) %>%
  mutate(
    year = as.integer(str_extract(date_added, "\\d{4}")),
    type = type
  )

# Analyze types of Turkish shows
turkish_types <- turkish_shows %>%
  count(type, sort = TRUE)

# Plot distribution of Turkish show types
ggplot(turkish_types, aes(x = "", y = n, fill = type)) +
  geom_bar(stat = "identity", width = 1) +
```

```
  coord_polar("y", start = 0) +
  labs(title = "Distribution of Turkish Content on Netflix",
       fill = "Type") +
  theme_void() +
  geom_text(aes(label = paste0(type, "\n(", n, ")")),
            position = position_stack(vjust = 0.5),
            size = 4)
```




```
# Analyze trends in Turkish content over time
turkish_trends <- turkish_shows %>%
  count(year, type) %>%
  complete(year, type, fill = list(n = 0))

# Plot trends in Turkish content
ggplot(turkish_trends, aes(x = year, y = n, color = type)) +
  geom_line() +
  geom_point() +
  labs(title = "Trends in Turkish Content on Netflix",
       x = "Year",
       y = "Number of Titles",
       color = "Type") +
  theme_minimal()
```



Most titles from my home country Turkey have been movies which makes sense as TV Shows are much harder to produce. There is a substantial decrease in the movies and TV Shows coming from Turkey in 2018 this might be caused by:

Regulatory and Political Issues: In 2018, Turkey introduced new regulations requiring streaming services to obtain licenses and comply with local broadcasting standards. This led to increased scrutiny and potential censorship, which might have affected the availability of Turkish content on platforms like Netflix1.

Content Strategy Shifts: Netflix often adjusts its content strategy based on viewer preferences and market trends. Around that time, Netflix might have shifted its focus to other regions or types of content that were gaining more popularity globally2.

Production Challenges: Political and economic instability in Turkey during that period could have impacted the production and export of TV shows and movies. This might have led to fewer Turkish productions being available for international distribution2.

Despite these challenges, Turkish content has seen a resurgence in recent years, with Netflix investing in new Turkish originals and licensing popular Turkish dramas3.

Reference: The data comes from the following TidyTuesday Github repository https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-04-20