

Anime EDA

Shelly

Load Packages

```
library(tidyverse)
```

Dataset Anime

Data source: TOP 1050 anime info was scraped from the MAL website in October 2025:
<https://www.kaggle.com/datasets/alexeluchian/mal-anime-dataset>

Brief description: Each observation represents an anime title (e.g., Attack on Titan) with its average viewer score, genre list, total episodes, brief synopsis, popularity rank, total member count, producing studio(s), and release year.

Research question 1: How do genres and studios influence anime ratings?

- Outcome variable (include the name/description and type of variable):

Name: score

- score represents the average viewer rating given to each anime. It reflects overall audience satisfaction and quality perception on a continuous scale from 0 to 10.
- It is influenced by characteristics such as the genre (e.g., action, romance, comedy) and studio (e.g., MAPPA, Kyoto Animation) responsible for production.

Type of variable: Continuous (numeric)

Research question 2: Is there a relationship between the number of episodes and viewer engagement?

- Outcome variable (include the name/description and type of variable):

Name: members

- members represents the total number of users who have added or interacted with an anime on the platform.
- It serves as a measure of viewer engagement — higher values indicate that more people are watching, following, or interested in the anime.

Type of variable: Continuous (numeric)

Load the data and provide a `glimpse()`:

```
library(tidyverse)
```

```
anime <- read.csv("https://github.com/lingyuehao/ids702_Team3/raw/refs/heads/main/anime_data.csv")
glimpse(anime)
```

```
Rows: 1,050
Columns: 9
$ title      <chr> "Attack on Titan", "Death Note", "Fullmetal Alchemist: Brot~
$ score      <dbl> 8.57, 8.62, 9.10, 8.48, 8.42, 7.22, 7.83, 9.03, 8.01, 7.79, ~
$ genres     <chr> "["Action", 'Award Winning', 'Drama', 'Suspense', 'Gore', '~
$ episodes   <dbl> 25, 37, 64, 12, 26, 25, 13, 148, 220, 12, 1, 12, 24, 24, 50~
$ synopsis   <chr> "Centuries ago, mankind was slaughtered to near extinction ~
$ popularity <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ members    <int> 4245518, 4186098, 3588803, 3443899, 3340293, 3240244, 32239~
$ studios     <chr> "["Wit Studio']", "["Madhouse']", "["Bones']", "["Madhouse'~
$ year       <int> 2013, 2006, 2009, 2015, 2019, 2012, 2016, 2011, 2002, 2014, ~
```

```
# Summary statistics
summary(anime)
```

```
# Check missing values
colSums(is.na(anime))
```

```
# Remove the 'synopsis' column
anime <- anime %>%
  select(-synopsis)

# Separate genre & Studio
anime_clean <- anime %>%
  mutate(genres = str_remove_all(genres, "\\[|\\]|'")) %>%
  separate_rows(genres, sep = ",\\s*") %>%
```

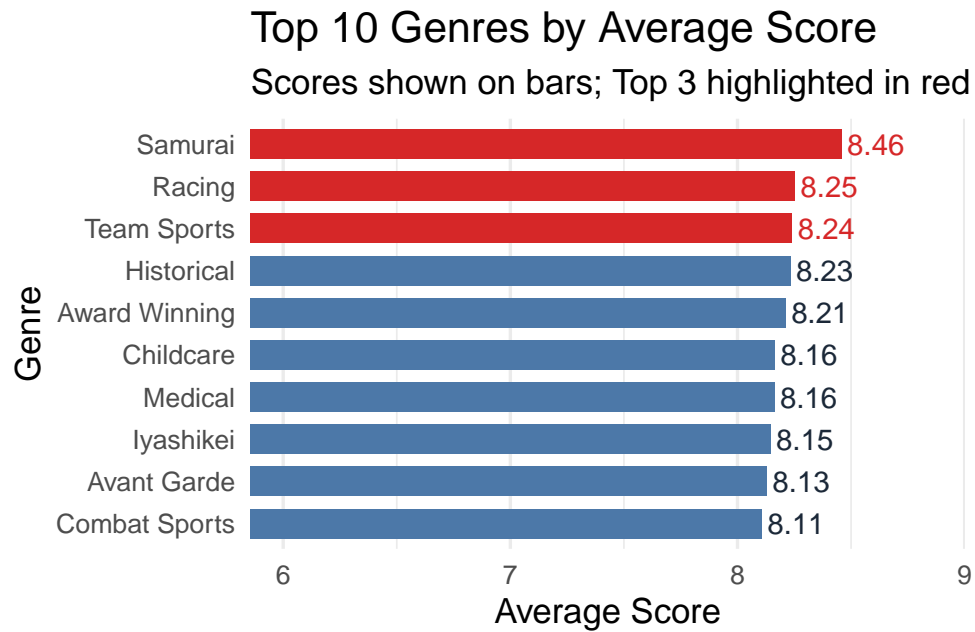
```

filter(!is.na(genres) & genres != "")

anime_clean <- anime_clean %>%
mutate(studios = str_remove_all(studios, "\\[|\\]|'")) %>%
  separate_rows(studios, sep = ",\\s*") %>%
  filter(!is.na(studios) & studios != "")

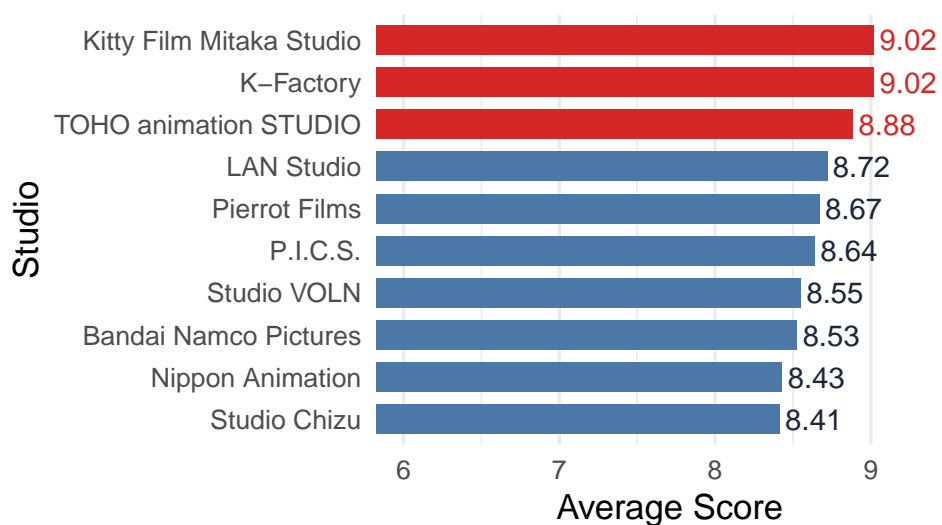
```

Exploratory Plots:



Top 10 Studios by Average Score

Scores shown on bars; Top 3 highlighted in



Average Members by Episode Range

