TidyTuesday_Pixar_Films

Yang Yuetong

2025-09-14

Suggested questions: Why are some values missing in the datasets? Which films have the highest score in each rating system? Are there distinct differences in ratings? Download the box_office dataset from the {pixarfilms} package. How does the box_office_us_canada value compare to the various ratings? Is the trend different for box_office_worldwide?

```
# Packages needed
# Data manipulation + tidying
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)

# Visualisation
library(ggplot2)

#sort
library(forcats)
```

```
# Import Pixar Films and Public Response Data from Github

pixar_films <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/
```

```
## Rows: 27 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): film, film_rating
## dbl  (2): number, run_time
## date (1): release_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
public_response <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/
```

```
## Rows: 24 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): film, cinema_score
## dbl (3): rotten_tomatoes, metacritic, critics_choice
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
box_office <- readr::read_csv('https://raw.githubusercontent.com/erictleung/pixarfilms/master/data-raw/
```

```
## Rows: 28 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): film
## dbl (4): budget, box_office_us_canada, box_office_other, box_office_worldwide
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Clean missing values to be recorded as NA
pixar_films <- pixar_films %>%
  mutate(
    film_rating = na_if(film_rating, "N/A"),
    film_rating = na_if(film_rating, "Not Rated")
)
```

```
#drop extra row index column
pixar_films <- pixar_films %>%
  select(-number)
```

```
#Find which column has missing values
colSums(is.na(pixar_films))
```

```
##         film release_date     run_time  film_rating
##            1            0            2            4
```

Upon inspecting the pixar_films dataset, film name is missing for the film that was released in 2023-06-16. A quick search of the release date and the run time would suggest that the film is Elemental. However, the run time is 103 minutes which does not match the data. run_time and film_rating also have missing values which could be due to inconsistent formatting.

```
colSums(is.na(public_response))
```

```
##            film rotten_tomatoes      metacritic    cinema_score  critics_choice
##               0               1               1               2               3
```

In the public_response dataset, the ratings for Luca is not available. We may drop the entire row. rating from cinema score is missing for Soul but other ratings from other critics could be useful.

```
public_response <- public_response %>% filter(film != "Luca")
```

Cinema score does not provide numerical scores, making it difficult to compare against other ratings.

```
public_response <- public_response %>% select(-cinema_score) #drop cinema score column
```

Convert the rating values to long format for data visualisation.

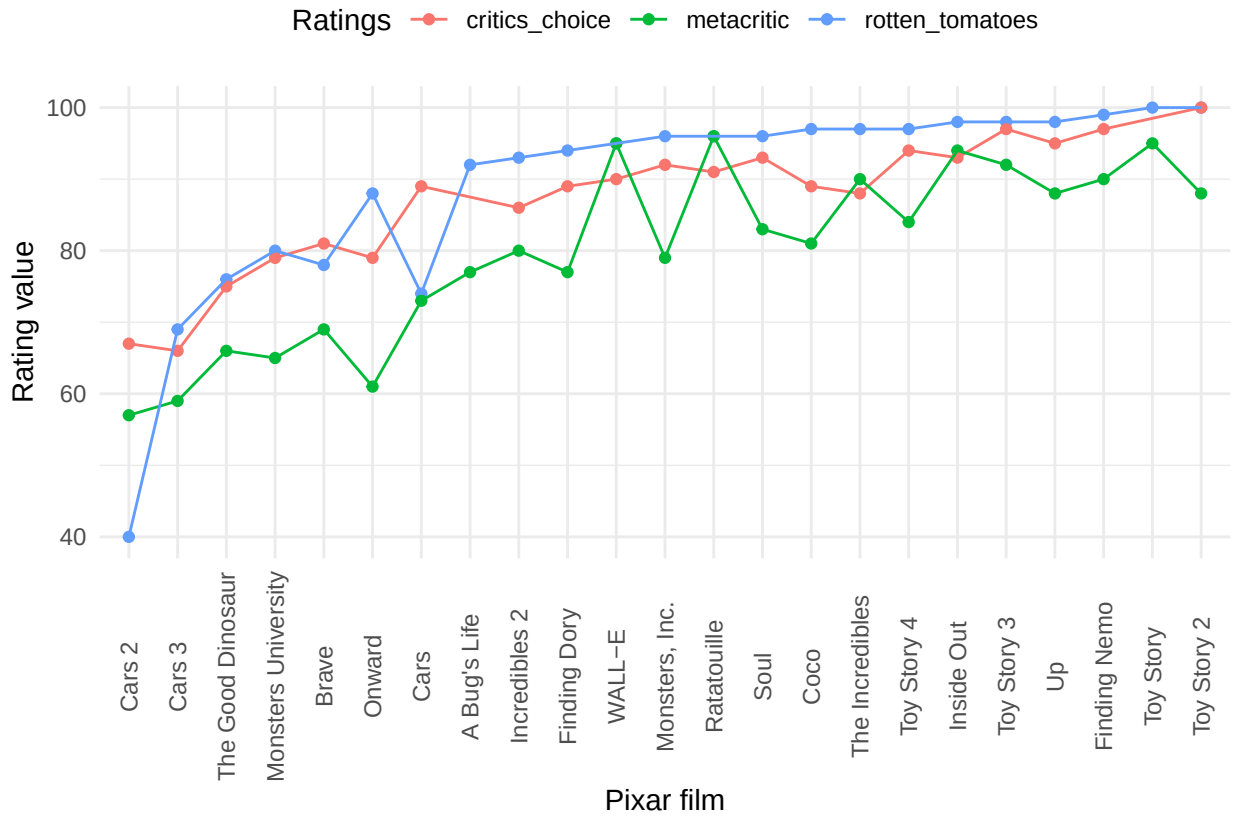
```
public_response <- public_response %>% pivot_longer(
  cols = c("rotten_tomatoes","metacritic","critics_choice"),
  names_to = "ratings",
  values_to = "ratings_value"
)
```

I want to drop rows where ratings is missing, then create a column containing max rating across different rating types, in order to sort the rows from the lowest to highest max rating.

```
public_response <- public_response %>%
  drop_na(ratings_value) %>%
  group_by(film) %>%
  mutate(max_rating = max(ratings_value, na.rm = FALSE)) %>%
  ungroup() %>%
  mutate(film = fct_reorder(film, max_rating, .desc = FALSE)) %>%
  arrange(film)
```

Plot a line graph to compare ratings for pixar films.

```
public_response %>%
  ggplot(aes(x = film, y = ratings_value, col = ratings, group = ratings)) +
  geom_point() +
  geom_line(aes(group = ratings)) +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "Pixar film", y = "Rating value") +
  guides(col = guide_legend(title = "Ratings")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5),
        legend.position = "top")
```



Firstly, rotten tomatoes has the highest rating for Toy Story 2 and Toy Story. Metacritic has the highest rating for Ratatouille. Critic choices has the highest rating for Toy Story 2. It seems like most people / critics like the older movies like Finding Nemo, Toy Story 1 and 2.

```
colSums(is.na(box_office))
```

```
##                  film                budget box_office_us_canada
##                     0                     1                     0
##     box_office_other box_office_worldwide
##                     0                     0
```

The missing value is in Luca's budget. If we are not using that column, we can remove it. We can also remove box_office_other.

```
box_office <- box_office %>% select(-budget,-box_office_other)
```

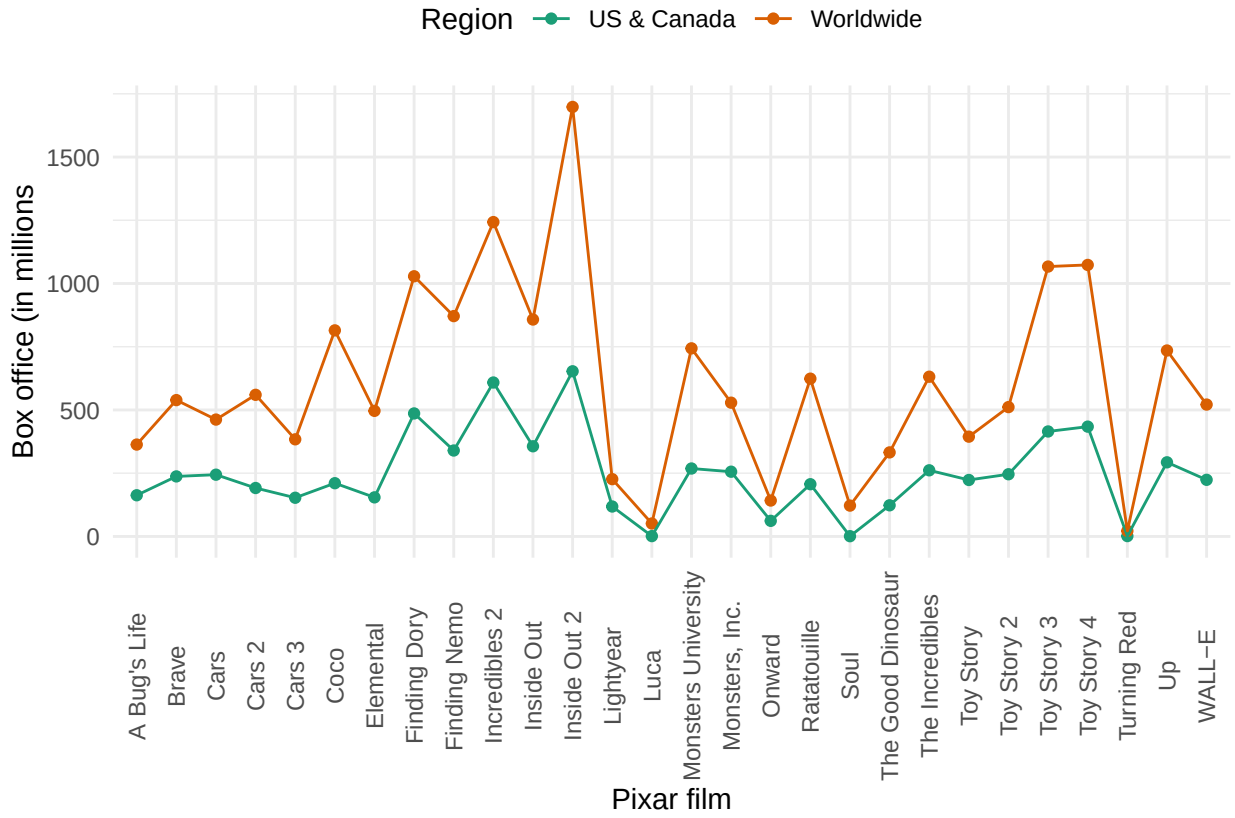
We will now compare the box_office in US/Canada vs Worldwide for all of the pixar_films

Similarly, convert the box_office values to long format for data visualisation.

```
box_office <- box_office %>% pivot_longer(
  cols = c("box_office_us_canada","box_office_worldwide"),
  names_to = "region",
  values_to = "box_office"
  ) %>%
  mutate(region = case_when(
```

```
    region == "box_office_us_canada" ~ "US & Canada",
    region == "box_office_worldwide" ~ "Worldwide",
    TRUE ~ region
  ))
```

```
box_office %>%
  ggplot(aes(x = film, y = box_office / 1e6, col = region, group = region)) +
  geom_point() +
  geom_line(aes(group = region)) +
  scale_color_brewer(palette = "Dark2") +
  labs(x = "Pixar film", y = "Box office (in millions") +
  guides(col = guide_legend(title = "Region")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5),
        legend.position = "top")
```



The trend seems to be similar for US/Cananda vs Worldwide. The largest difference is for Inside Out 2 suggesting that alot more people outside US/Canada watch Inside Out 2, compared to other pixar films.