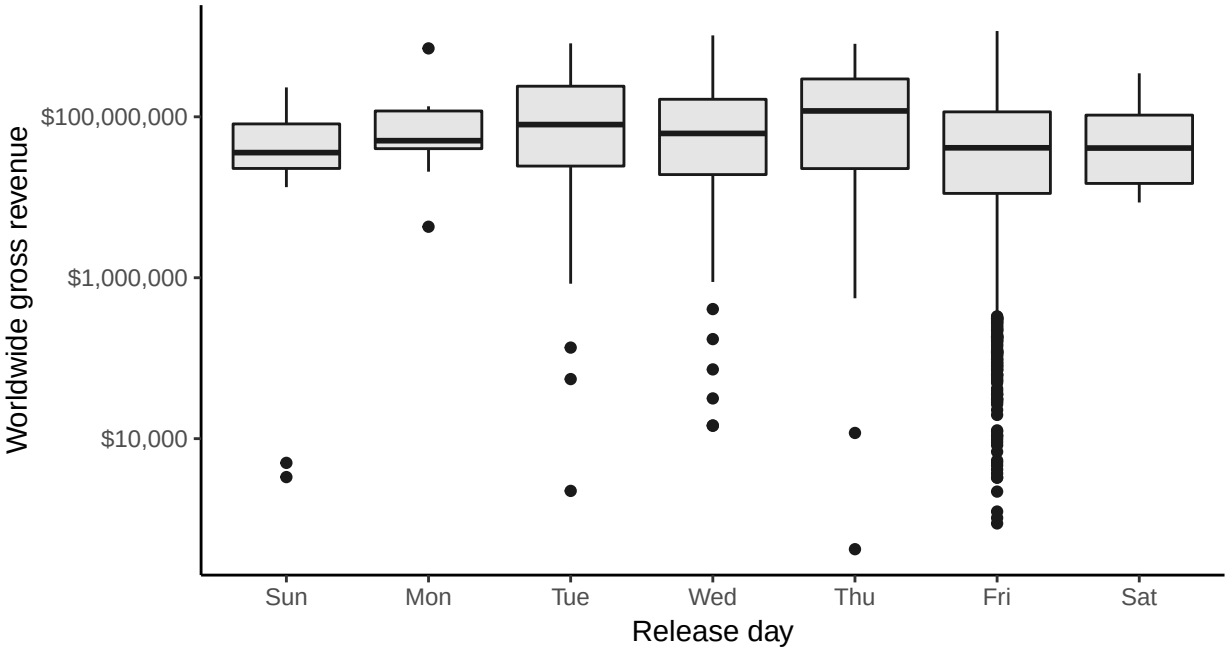
Refresher: Tidy Exploratory Data Analysis

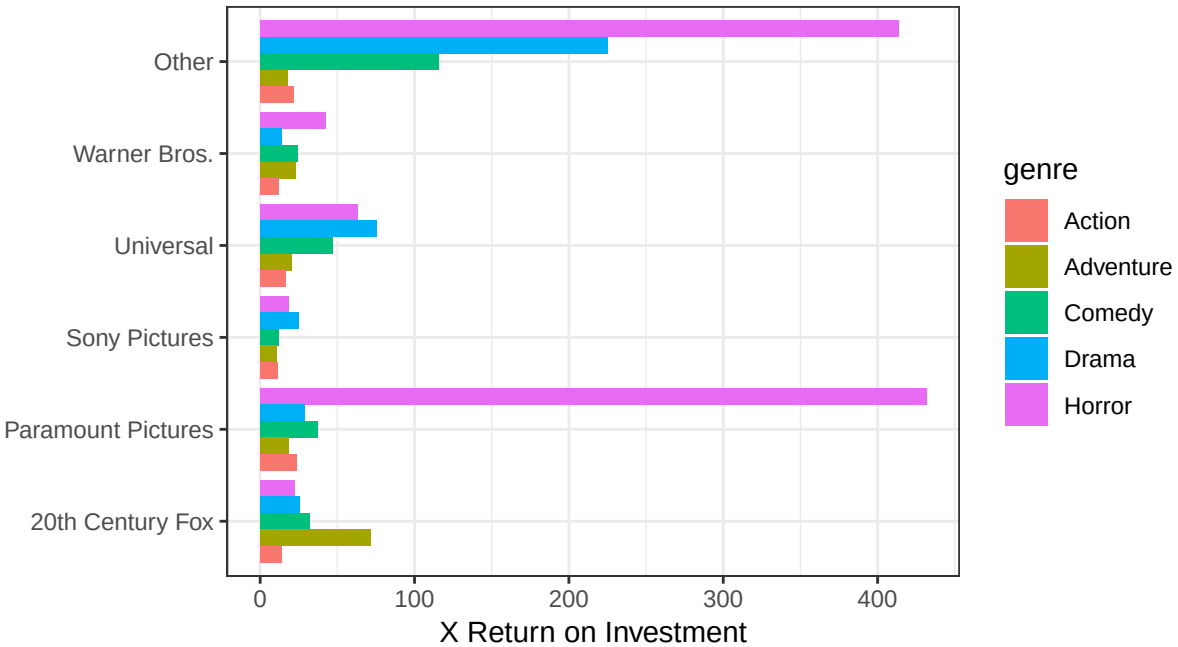
Exercise #1

Does the day a movie is release affect revenue? Make a boxplot showing the worldwide revenue for each day.



Exercise #2

What does the ROI look like for each distributor by genre? Modify the code used for worldwide revenue.



Exercise #3

Join the movies data to the imdb ratings.

```
imdb <- read_csv("data/movies_imdb.csv")
movimdb <- inner_join(mov, imdb, by="movie")
```

1. Separately for each MPAA rating, display the mean IMDB rating and mean number of votes cast.
2. Do the same but for each movie genre.
3. Do the same but for each distributor, after lumping distributors in a mutate statement to the top 4 distributors, as we've done before.
4. Create a boxplot visually summarizing what you saw in #1 and #2 above. That is, show the distribution of IMDB ratings for each genre, but map the fill aesthetic for the boxplot onto the MPAA rating. Here we can see that Dramas tend to get a higher IMDB rating overall. Across most categories R rated movies fare better. We also see from this that there are no Action or Horror movies rated G (understandably!). In fact, after this I actually wanted to see what the "Horror" movies were having a PG rating that seemed to do better than PG-13 or R rated Horror movies.
5. Create a scatter plot of worldwide gross revenue by IMDB rating, with the gross revenue on a log scale. Color the points by genre. Add a trendline with `method="lm"`.
6. Create the same plot, this time putting the number of votes on the x-axis, and make both the x and y-axes log scale.
7. Create the above plots, but this time plot the ROI instead of the gross revenue.
8. Is there a relationship between the release date and the IMDB ratings or votes cast? Surprisingly, there doesn't appear to be one.
9. Is there a relationship between the IMDB rating and the number of votes cast? It appears so, at least as you get toward the movies with the very largest number of ratings.
10. Looking at that above plot, I'm interested in (a) what are those movies with the largest number of votes? and (b) what are those movies with at least 50,000 votes that have the worst scores?

Exercise #4

Read in, clean up the `grads.csv` data.

```
grads_raw <- read_csv(here::here("data", "grads.csv"), col_types=cols())
grads <- grads_raw %>%
  arrange(desc(Median)) %>%
  mutate(Major = str_to_title(Major)) %>%
  mutate(Major = fct_reorder(Major, Median)) %>%
  mutate(Major_category = fct_reorder(Major_category, Median)) %>%
  mutate(pct_college=College_jobs/(College_jobs+Non_college_jobs)) %>%
  filter(!is.na(pct_college)) %>%
  filter(!is.na(Total))
```

Remake table 1 from the FiveThirtyEight article.[1]

- Use the `select()` function to get only the columns you care about.
- Use `head(10)` or `tail(10)` to show the first or last few rows.

[1] https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/

Exercise #5

What about "underemployment?" Which majors have more students finding jobs requiring college degrees? Make a boxplot of each major category's percentage of majors having jobs requiring a college degree (`pct_college`). Reorder the `Major_category` with `fct_reorder()` to order it by the `pct_college` variable we created, before plotting it. Flip the axis and expand the y-axis to include zero.