Week 5 Day 1 Activity: Revisiting IMDb Movie Reviews — Comparing Two Groups

This week our focus is comparing the means of two groups. With linear regression we were able to compare the predicted mean response across different values of a continuous explanatory variable. This week, however, we are moving from a *continuous* explanatory variable to a **categorical** explanatory variable!

Movies released in 2016

Recall the dataset we used a couple of weeks ago, visualizing the distribution of IMDB movie ratings. The dataset included eight variables, listed below.

Variable	Description
`movie_title`	Title of movie
`budget_mil`	Amount of money (in US $ millions) budgeted for the production of the movie
`revenue_mil`	Amount of money (in US $ millions) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`facebook_likes`	Number of likes a movie receives on Facebook

Comparing Two Groups

Let's compare the budget for movies rated PG and movies rated R. Technically, the dataset has four different content ratings: Not Rated, PG, PG-13, and R. To make the comparison of PG and R movies easier, I went ahead and made a smaller dataset containing only these two ratings.

In case you are interested, this is the code I used!

```
movies_pg <- filter(movies,
                    content_rating %in% c("PG", "R")
                    )
```

Similar to summarizing the mean of all of the movies, we have two options to compare these two groups:

1. Use summary statistics
2. Use visualizations

Summary Statistics

Let's start with summary statistics. Our familiar friend `favstats()` can help us compare summary statistics across different groups. Before when we used `favstats()` we only had one variable, but now we have two!

Now, we will use two variables as a "formula", which looks like `response ~ explanatory`. So, the budget of the film is the response and the content rating is the explanatory variable. So, our code looks like:

```
favstats(budget_mil ~ content_rating,
         data = movies_pg)
```

Use the output from the `favstats()` function to answer the following questions:

```
##   content_rating min    Q1 median      Q3 max     mean       sd  n missing
## 1             PG 0.5 11.00   74.0 151.250 175 86.54167 71.52795 12       0
## 2              R 0.0  7.75   19.5  29.625  60 21.09375 16.99926 32       0
```

1. Report the mean budget amount for the PG rating. Use appropriate notation.

2. Report the mean budget amount for the R rating. Use appropriate notation.

3. Calculate the difference in mean budget amount for movies in 2016 with a PG rating minus those with a R rating. Use appropriate notation with informative subscripts.

4. Which content rating has the largest standard deviation?

5. Which content rating has the largest sample size? What effect does sample size have on each group's standard deviation?

6. Which content rating has the largest IQR?

7. Based on the summary statistics, roughly sketch what you believe the distribution of budgets for each content rating would look like.

Visualizing Differences

Let's refresh ourselves on the different ways to plot a numerical variable.

8. What are the three types of plots used to plot a single quantitative variable?

9. For each type of plot, how would you include a categorical variable in the plot?

Faceted Histograms

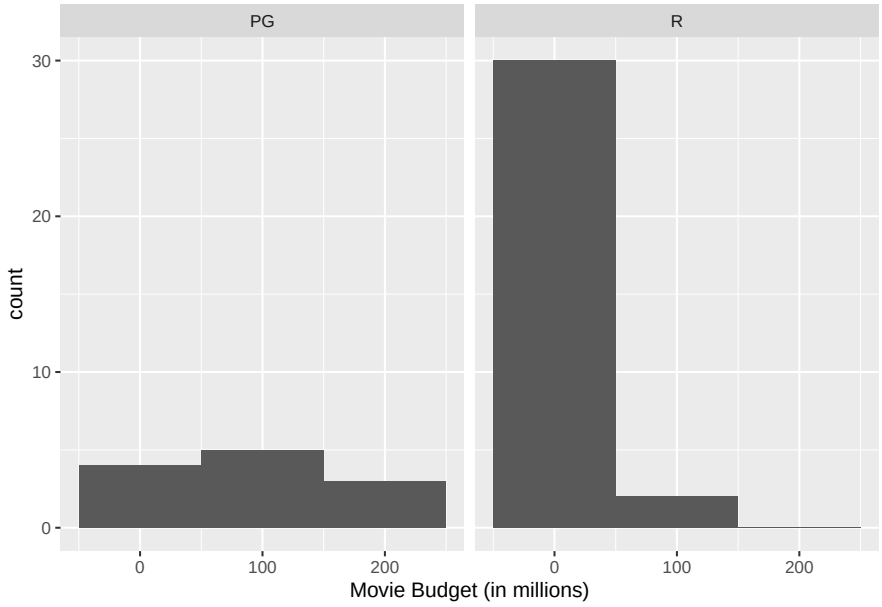
When we want to add a categorical variable (like `content_rating`) to a histogram, we create separate plots for each level of the categorical variable. These separate plots are called **facets**. We are comparing the budgets for PG and R movies, so we will have two facets, one per content rating.

The code to make a faceted histogram looks like the following:

```
ggplot(data = movies_pg,
       mapping = aes(x = budget_mil)) +
  geom_histogram(binwidth = 100) +
  labs(x = "Movie Budget (in millions)") +
  facet_wrap(~ content_rating)
```

Notice the last line is the only new part! That line creates a faceted plot (using `facet_wrap()`) and says to facet "by" (`~`) the `content_rating`.

Let's look at what this plot ends up looking like:



10. How would you describe the shape of the distribution of PG movie budgets?

11. How would you describe the shape of the distribution of R movie budgets?

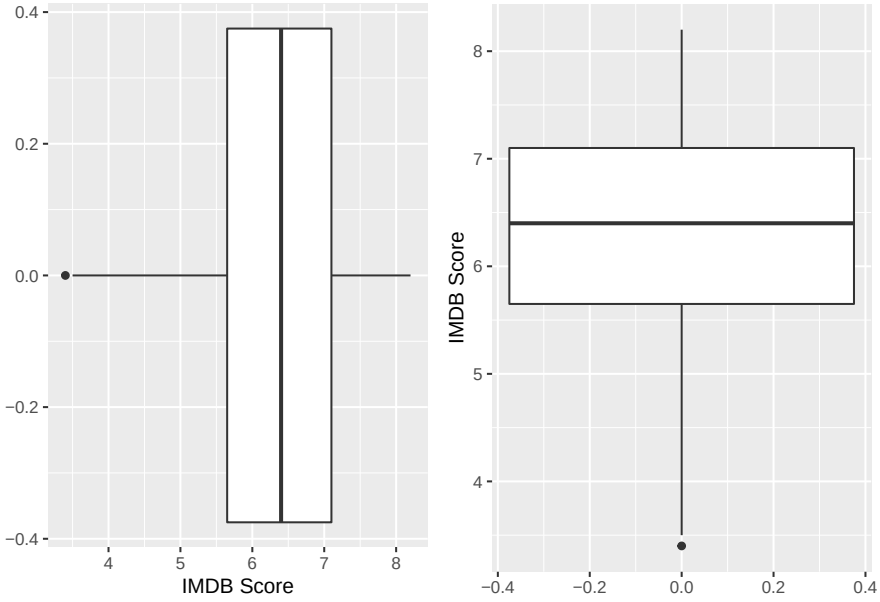
12. Visually, does it appear that there are outliers in either of these groups?

13. How do these distributions compare to what you expected in your sketches (in #7)?

Side-by-Side Boxplots

Another way we can incorporate a categorical into our plots is to plot our boxplots for each group side-by-side. As opposed to faceting, these boxplots will be on the **same** plot. We only need to add one extra piece to our previous code: a categorical variable.

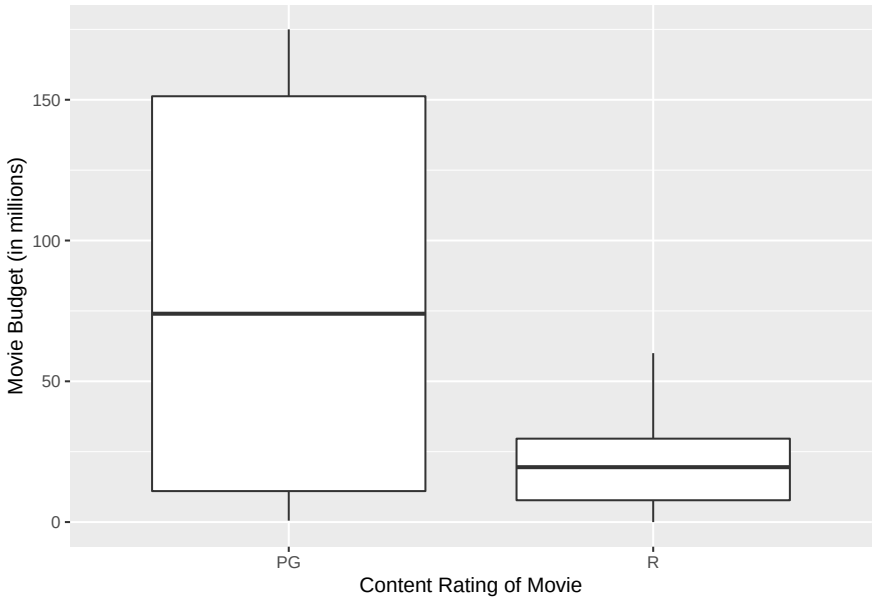
Before, we either plotted our **one** numerical variable horizontally (using `x`) or vertically (using `y`).



Now. we need to plot **two** boxplots side-by-side. Similar to before, we can stack the plots horizontally or vertically.

Horizontal Stacking

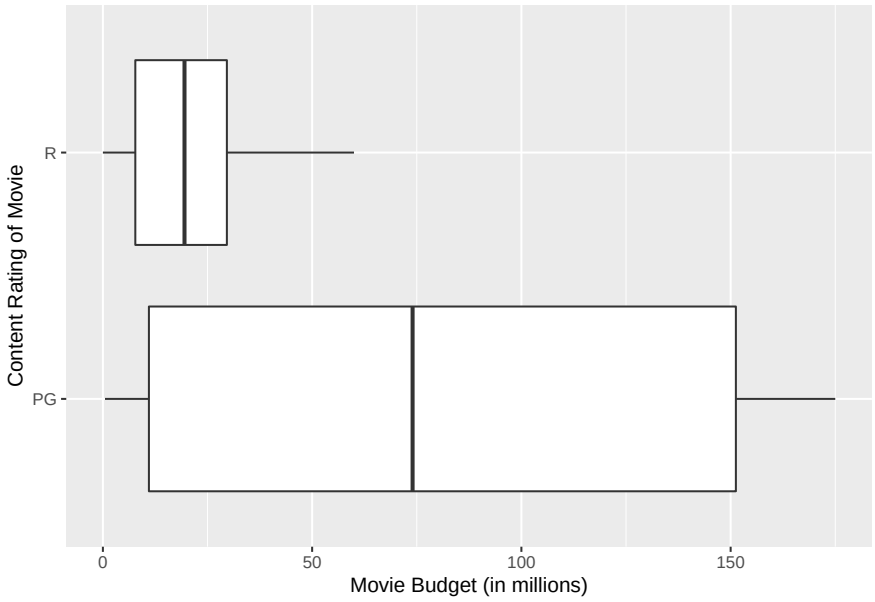
```
ggplot(data = movies_pg,
       mapping = aes(y = budget_mil, x = content_rating)) +
  geom_boxplot() +
  labs(x = "Content Rating of Movie",
       y = "Movie Budget (in millions)")
```



14. Why are the boxplots stacked side-by-side horizontally? What part of the `R` code does this?

Vertical Stacking

```
ggplot(data = movies_pg,
       mapping = aes(x = budget_mil, y = content_rating)) +
  geom_boxplot() +
  labs(y = "Content Rating of Movie",
       x = "Movie Budget (in millions)")
```



15. How was the previous code changed to stack the boxplots side-by-side vertically?

16. Which orientation do you prefer?