# Homework 8

**Due: Wednesday April 27 at 7pm**

Use RStudio for all exercises. You should provide one script that contains all the code and includes comments noting which code is for which exercises. You will also need to comment on the results, so place them in a Word (or Open Office or HTML or PDF) document and add your comments to answer the questions. If you are familiar with knitr and would like to use that to programmatically create your document, that would be fine.

Any code based on code from elsewhere (e.g. code provided with the text) must reference in comments the source of the original code. **Script files should be the actual script files**, not code pasted into some other document. You will need to submit one script file and one report file. You do not need to submit them as a zip file.

All exercises are based on the IMDB **movies** data set (**ggplot2movies::movies**) from the **ggplot2movies** package for R. Some initial data definitions are provided in **HW8Setup.R** in the Homework 8 directory on compass. It contains a data set called **newdata** which contains the original movies data, but rounds the **rating** variable so the ratings are now integers 1 through 10. The **mpaaratings** data set contains integer **rating** values, **mpaa** values and counts for those categories. Rating is the numeric average of ratings given by IMDB users for a movie. mpaa is the rating (PG, R, etc.) given to a movie. Note that not all movies in IMDB have MPAA ratings listed, and older movies would not have had MPAA ratings at all.

## Exercises for All Students

### Exercise 1:

Use the **mpaaratings** data set and treemaps to analyze the following items

- relative proportions of IMDB rating values in the data
- relative proportions of MPAA rating values in the data
- proportion of IMDB ratings within each MPAA rating group
- proportion of MPAA rating group for each IMDB rating value

For each of the treemaps, comment on noticeable differences across groups and what that tells us about IMDB ratings and MPAA ratings. For instance, which IMDB ratings are more common and less common? Which MPAA ratings are more common or less common in general? How do the proportions of IMDB ratings compare across MPAA ratings? And how do MPAA ratings compare across IMDB rating groups?

(Hint for plots: sometimes it is clearest to just use counts for the size and the color of the boxes.)

### Exercise 2:

For this exercise, you will need to create new data sets from the **newdata** data set defined in **HW8Setup.R**.

a) Create a streamgraph for rounded IMDB ratings counts over time (e.g. it should show the number of movies with each IMDB rating for each year) and comment on any interesting trends. You should note anything that seems to be consistent over time and anything that seems to have changed over time.
b) Create a streamgraph for MPAA ratings counts over time (e.g. the number of movies with each MPAA rating each year) and comment on any interesting trends. You should note anything that seems to be consistent over time and anything that seems to have changed over time.

## Additional Exercise for Graduate Students

## Exercise 3:

Start with the original **movies** data.

a) Create a new data set called **kdedata** that contains the rating and length variables for all movies in the original **movies** data set that have length 180 or less (so the **kdedata** data set will contain the IMDB ratings and movie lengths for all movies that are 3 hours long or shorter). Make sure to start with the **movies** data set so we have the actual **rating** values and not the rounded values.
b) Create a two dimensional kernel density estimate plot for the **rating** and **length** values in **kdedata**. Adjust the bandwidths as you see fit to obtain a reasonably smoothed estimate of the distribution of rating and length values. Interpret what the density estimation tells us about more likely lengths and ratings for movies and any apparent relation between lengths and IMDB ratings of movies.