# Data Visualization using R Workshop - Homework

*Chester Ismay*

*September 30, 2015*

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: (`cars` is a data set automatically loaded into R.)

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

### Load in the data

The problems here will have you examine the `movies` dataset included in the `ggplot2` package.

```r
data(movies, package = "ggplot2")
```

I recommend entering `?movies` into the Console to get an idea of how the data was collected and what the columns correspond to in the data set.

Now that the data is loaded try the following problems. Remember that you will need to include all of your R code in Chunks. I've added a few blank chunks below to get you started. Note that I've also labeled the chunks corresponding to their problem number. It's always a good habit to label your chunks!

### Problems

1. Is there a relationship between longer movies and the number of votes the movies received on IMDB?

2. Same question as Problem 1 but use only movies with `length` less than 240 minutes. Also use the `alpha` parameter to `geom_point` to deal with over-plotting.

3. Use the `geom_smooth` function to add a regression line (without confidence bands) to the plot in Problem 2.

4. What does the distribution of movies in regards to MPAA rating (by count) look like for this dataset? (Exclude those without an MPAA rating.) Use a light color for your plot. A list of all built-in R colors is at http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf.

5. How does the `r10` variable relate to overall rating? Additionally, look into whether or not the movie is a documentary. (Note that if you are trying to color based on the values of a variable you will need to enclose it in the `aes()` function.) *Hint*: You may find it useful to wrap the genre columns in `factor()` to convert them to a categorical variable instead of a numerical. Check to see what happens when you don't include `factor()` below.