Homework3a - Web Scraping and dplyr

Problem 1

IMDB has a list of the top 250 movies at this URL.

```
  url <- "http://www.imdb.com/chart/top"
```

Use whatever method of web scraping you think is appropriate to get all the names of the top 250 movies, their ratings, and their year of release.

Problem 2

Given the data set from #1:

- What years produced more of the top 250 movies?
- Give the average of all the ratings, and the average per 5 year span. Show any intermediate work.
- Using the examples from the slides and a little research, give a histogram of the movies released, with bucket size of 5 years.

Use the Open Movie Database API (http://www.omdbapi.com/) JSON API to query more information about the top movies. You will want to read the site's page for the API usage. One example URL when searching for Shawshank by movie name is `http://www.omdbapi.com/?t=the+shawshank+redemption&y=1994&type=movie` , which returns:

```
{
    Title: "The Shawshank Redemption",
    Year: "1994",
    Rated: "R",
    Released: "14 Oct 1994",
    ...
    Metascore: "80",
    imdbRating: "9.3",
    imdbVotes: "1,521,105",
    imdbID: "tt0111161",
    Type: "movie",
    Response: "True"
}
```

Problem 4

Query your favorite movie from OMDB, and print the name, plot, and awards.

Problem 5

Query any of set of 40 movies from the top 250 movies listed in problem #1. Build a data frame with all data that comes back from the queries. Be sure to include the Rotten Tomatoes Ratings. If you want a challenge, try to get all 250 movies.

In order to prevent flooding this free API service, you need to pause between each couple of queries for a little bit. Use the `Sys.sleep` function to do this, as shown in the example below:

```
testit <- function(x)
{
    p1 <- proc.time()
    Sys.sleep(x)   # nothing happens for x seconds
    proc.time() - p1
}
testit(3.7)
```

- Hint 1: Use the examples on the main web page or your browser to test searching for 1 URL.
- Hint 2: Write some code that works for 1 movie query, then build the right iteration over that function.
- Hint 3: In either case, once you download and clean all the data, it will help to save it to your harddrive.

Problem 6

Give the summary statistics for the ‘Metascore’, ‘imdbRating’, and ‘imdbVotes’ across the 40 movie sample set. Give the same break down, but by movie age rating (i.e. PG-13, R, etc).

Problem 7

Give the summary statistics for all the numeric Rotten Tomato rating values, and compare the overall ratings against the IMDB ratings.

- Give the same statistics for broken down by movie age rating.
- Do the same for each 5 or 10 year span.

Problem 8

Are there any other interesting patterns you can find in this data set?