Project 2 - Rotten or fresh?

Rotten Tomatoes is a film review aggregator website devoted to reviews, information, and news of films. Its name is derived from the cliché of audiences throwing rotten tomatoes or vegetables at a poor stage performance. Find out more about how the website works at http://en.wikipedia.org/wiki/Rotten_Tomatoes.

In this project you will be predicting audience scores (a measure of how much the public likes a movie) using a variety of information on movies from Rotten Totamoes, IMDB, Oscar wins and nominations, etc. The dataset is comprised of 426 randomly sampled movies released between 1970 and 2012.

You can download the dataset (called `movies`) here.

```
download.file("http://stat.duke.edu/courses/Fall13/sta101/projects/movies.Rdata", destfile = "movies.Rdata")
load("movies.Rdata")
```

Below is a description of the variables:

1. `title`: Title of movie
2. *`audience_score`: Audience score on Rotten Tomatoes (response variable)
3. `type`: Type of movie (Documentary, Feature Film, TV Movie)
4. `genre`: Genre of movie (Action & Adventure, Comedy, Documentary, Drama, Horror, Mystery & Suspense, Other)
5. `runtime`: Runtime of movie (in minutes)
6. `year`: Year the movie is released
7. `mpaa_rating`: MPAA rating of the movie (G, PG, PG-13, R, Unrated)
8. `imdb_num_votes`: Number of votes on IMDB
9. `critics_score`: Critics score on Rotten Tomatoes
10. `critics_rating`: Categorical variable for critics rating on Rotten Tomatoes (Certified Fresh, Fresh, Rotten)
11. `best_pic_nom`: Whether or not the movie was nominated for a best picture Oscar (no, yes)
12. `best_pic_win`: Whether or not the movie won a best picture Oscar (no, yes)
13. `best_actor_win`: Whether or not one of the main actors in the movie ever won an Oscar (no, yes) – not that this is not necessarily whether the actor won an Oscar for their role in the given movie
14. `best_actress_win`: Whether or not one of the main actresses in the movie ever won an Oscar (no, yes) – not that this is not necessarily whether the actresses won an Oscar for their role in the given movie
15. `best_dir_win`: Whether or not the director of the movie ever won an Oscar (no, yes) – not that this is not necessarily whether the director won an Oscar for the given movie
16. `top200_box`: Whether or not the movie appears in the Top 200 Box Office listing (from: http://boxofficemojo.com/alltime/adj

    The following variables are included as additional information, but should not be used in the analysis
17. `audience_rating`: Categorical variable for audience rating on Rotten Tomatoes (Spilled, Upright)
18. `director`: Director of the movie

19-23. `actor1-actor5`: List of first 5 main actors in the movie (abridged cast), this information was used to determine whether the movie casts an actor or actress who won a best actor or actress Oscar

24. `rt_url`: Link to Rotten Tomatoes page for the movie
25. `imdb_url`: Link to IMDB page for the movie
26. `imdb_id`: IMDB ID of the movie

You are welcomed to use the data set as is, or manipulate the variables (collapsing levels, etc.), or create new variables as you see fit.

Poster & presentation

Your results will be displayed on a poster during your scheduled lab time on Monday, December 2. Each team will have 4 minutes scheduled during which all team members should be at the poster to give a formal presentation. All team members must contribute to this presentation. Note that this is too short a time to go through in detail all of your analysis and findings. Your poster should be showcasing those. The presentation should only highlight what you think are most interesting findings, or something that you believe may be unique about your approach.

For the rest of the poster session, you will be wandering around, assessing the other teams' posters and asking questions. The lab time will be broken up so each person will spend 1/(# of members in team) of the remaining time by your own poster to answer questions. This is part of your grade, so every team member should feel comfortable answering questions regarding all aspects of the project.

Paper

Maximum of 10 pages (including figures, etc.). Due by Thursday, December 5, at the beginning of class. Hard copy + submission on Sakai (.Rmd file). Be conservative in what you include in your paper, 10 pages is not very long.

```
download.file("http://stat.duke.edu/courses/Fall13/sta101/projects/prj2.Rmd", destfile = "prj2.Rmd")
```

Components of project

You can focus on whatever aspects of the data you find interesting, but as a minimum you must include each of the following components:

1. **Introduction:** Explain your data, including implications for the scope of inference.
2. **Univariate analysis:** Analysis of variables of interest, one at a time
3. **Bivariate analysis:** Exploring and assessing relationships between two variables at a time (response vs. explanatory or explanatory vs. explanatory) using appropriate methods
4. **Multiple regression:**
   (a) Decide on a "best" model for predicting the response variable. You do not need to explain every step you took to arrive at this model, but should give some indication of why you chose the model you did.
   (b) For your best model, are the conditions met? If not, what are the implications?
   (c) Using your best model, obtain a predicted value (and the associated confidence interval) for the audience score for a 2013 movie of your choosing. (You will need to find relevant information about the movie online.)
5. **Conclusion:** (See below.)

After providing the description of your dataset and research question in the introduction you must apply what you have learned about descriptive statistics, graphical methods, correlation and regression, and hypothesis testing and confidence intervals to your dataset. The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather showcase that you are proficient at using R at a basic level, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research question. Also pay attention to your presentation. Neatness, coherency, and clarity will count.

Your write up must also include a roughly one page conclusion and discussion. This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. Also critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the statistical analysis should be discussed here. A paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project should also be included.

Grading

Grading of the project by the professor and TAs will take into account the following:

- Content - What is the quality of research and/or policy question and relevancy of data to those questions?
- Correctness - Are statistical procedures carried out and explained correctly?
- Writing and Presentation - What is the quality of the statistical presentation, writing and explanations?
- Creativity and Critical Thought - Is the project carefully thought out? Are the limitations carefully considered? Does it appear that time and effort went into the planning and implementation of the project?

Your grade for the project will be based on the following components:

- 40% Poster (poster, presentation, answers to questions)
- 40% Paper
- 10% Classmates' grades (based on poster session)
- 10% Team member evaluation

Team member evaluation: All of the other components are common to the entire team. This is your chance to be rewarded (or penalized) for contributing (or not contributing) to the team. You will rate each of your team members' contribution to the project on a scale of 0 - 5. You will receive an email with a link to a Qualtrics poll to submit your evaluations, these must be submitted by Friday, December 6, at 5pm. These ratings are anonymous to your teammates – please be honest. The average of these scores will be your score for 10% of your project grade. For grades less than 3, please provide some explanation. If any individual gets an average peer grade less than 1, this person will receive half the grade of the rest of the team.

A general breakdown of scoring is as follows:

> 90%-100% - Outstanding effort. Student understands how to apply all statistical concepts, can put the results into a cogent argument, can identify weaknesses in the argument, and can clearly communicate the results to others.
>
> 80%-89% - Good effort. Student understands most of the concepts, puts together an adequate argument, identifies some weaknesses of their argument, and communicates most results clearly to others.
>
> 70%-79% - Passing effort. Student has misunderstanding of concepts in several areas, has some trouble putting results together in a cogent argument, and communication of results is sometimes unclear.
>
> 60%-69% - Struggling effort. Student is making some effort, but has misunderstanding of many concepts and is unable to put together a cogent argument. Communication of results is unclear.
>
> Below 60% - Student is not making a sufficient effort.

Honor Code:

You may not discuss this project in any way with anyone outside your team, besides the professor and TAs. Failure to abide by this policy will result in a 0 for all teams involved.