STAT 6430 **Project**

This project centers on several related data sets, with the centerpiece the set `reviews.txt` that consists of 100K movie reviews. There are 1682 movies reviewed by 943 reviewers, with each rating an integer between 1 (lowest) and 5 (highest). Other date files include information about each reviewer, about each movie, and relating zip codes to states/territories. The file `README.txt` gives more information about the data sets.

The main goal of this project is try to answer two related questions:

1. What seems to be associated with a high rating?
2. What groups are most likely to provide higher ratings?

These questions might have related answers. Please confine your work to the given data sets, but otherwise free to take your investigation in any direction you like. Your work should be summarized in a written report of no more then 3 pages. You are encouraged to use graphs in addition to text to describe your findings.

In addition to the main investigation of your project, your written report should include an appendix (not counted in the 3 page limit) that answers the following questions:

1. Which percentage of each rating was given?
2. Which reviewers were the top-10 in terms of number of movies reviewed? (Provide the reviewer number and the number of movies reviewed. If there is a tie for 10th place, include all that tied.)
3. Find a 95% confidence interval for the average rating among all reviewers, and a 95% confidence interval for the average rating among the top-10 reviewers. Does there appear to be evidence that the two groups differs?
4. Which movies were the top-10 based on of number of times reviewed? (Provide the movie title and the number of times reviewed. If there is a tie for 10th place, include all that tied.)
5. Which genre occurred most often, based on the number of reviews. Which was least often? (Don't include "unknown" as a genre for this question.)
6. What percentage of reviews involved movies classified in at least two genres?
7. Give a 95% confidence interval for the average rating for male reviewers, and do the same for female reviewers.
8. Which state/territory/Canada/unknown produced the top-5 most reviews?
9. What percentage of movies have exactly 1 review? 2 reviews? 3 reviews? Continue to 20 reviews.
10. Which genre had the highest average review, and which had the lowest average review?
11. Repeat the previous question, for reviewers age 30 and under and then for reviewers over 30.

Besides your written report, your team will also give a 8-minute (maximum!!) presentation describing your results. Please bring a printed copy of your report and presentation slides to your presentation.

Other information:

- You should upload electronic versions of your R code, report, and presentation slides.
- Your presentation should be in Powerpoint or a variant. I don't recommend that you plan to connect your laptop to the projector or use online resources, but how you use your 8 minutes is up to you.
- Static output from R is likely more reliable than getting R to work in front of a group.
- The presentations will be on Wednesday, August 2, 9:00-11:30am. (The written report is due then.) The team assignments and order are:

  6, 8, 7, 9, 11, 1, 13, (break) 12, 10, 3, 5, 2, 4

- The project grade will be based on a combination of the R code, the written report, and your presentation. Each team member will share the same project grade.