STAT 3250 Assignment 9

Directions: Please submit your Python code file, formatted as in previous assignments. You will also be uploading the output file described below.

The grader should be able to run your code when it is placed in the same directory as the input data file. Be sure that your code loads any libraries you are using.

Background: This project centers on several related data sets, the centerpiece the set `reviews.txt` which consists of 100K movie reviews. There are 1682 different movies reviewed, with a total of 943 different reviewers. Each review is an integer between 1 (lowest) and 5 (highest). Other data files include information about each reviewer, about each movie, and relating zip codes to states/territories. The file `README_assign09.txt` gives more information about the data sets, including some that is required to answer the questions correctly. (So read it!)

Please confine your work to the given data sets. (Work with the zip code file given – part of the assignment is data cleaning.) Use these data sets to answer the following questions.

1. Find the 5 reviewers with the most reviews, and then use their reviews to find a 95% confidence interval for their average rating (taken as a group). Then find the average rating for the remainder of the reviewers. Is this average within the top-5 confidence interval? Here the sample sizes are quite large, so we can use the confidence interval formula

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

where s is the standard deviation with `ddof = 1`.

2. Which movies were the top-10 based on of number of times reviewed? (Provide the movie title and the number of times reviewed for each. If there is a tie for 10th place, include all that tied.)

3. Which genre occurred most often, based on the number of reviews. Which was least often? (Don't include "unknown" as a genre.)

4. What percentage of reviews are for movies classified in at least two genres?

5. Give a 95% confidence interval for the average rating for male reviewers, and do the same for female reviewers.

6. Which locations (state, territory, or Canada) formed the top-10 for number of reviews? (Provide a table of location and number of reviews. The location 'unknown' should not be included.)

7. Find the occupations that gave the highest average reviews, and the lowest average reviews. (Here "other" and "none" are not occupations, but "student" is.)

8. What percentage of movies have exactly 1 review? 2 reviews? 3 reviews? Continue to 20 reviews.

9. Which genre had the highest average review, and which had the lowest average review?

10. Suppose that a "positive review" is one with a rating of 4 or 5.

    (a) Find a 95% confidence interval for $p_f - p_m$, where p_f is the proportion of positive reviews from females and p_m is the proportion of positive reviews from males. Is there evidence that the proportions differ?

    (b) It is thought that Canadians are nicer than Americans. Find a 95% confidence interval for $p_C - p_A$, where p_C is the proportion of positive reviews from Canadians, and p_A is the proportion of positive reviews from Americans. (Exclude those whose location is unknown.) Is there evidence that Canadians give more positive reviews?