BSAN440 (Fall 2019) Assignment 2

Shaobo Li

Due on 9/23/2019

Instruction

You may work with partner(s) on this assignment, but you are strongly recommended to solve each question by yourself before or after your group meetings. **Please note, each group will only submit one copy through Blackboard. Late submission will not be accepted.**

Please copy the sentence below to the first page of your submission document, and sign.

By signing below, I certify that all results are original and produced by members in our group.

Name:_____________

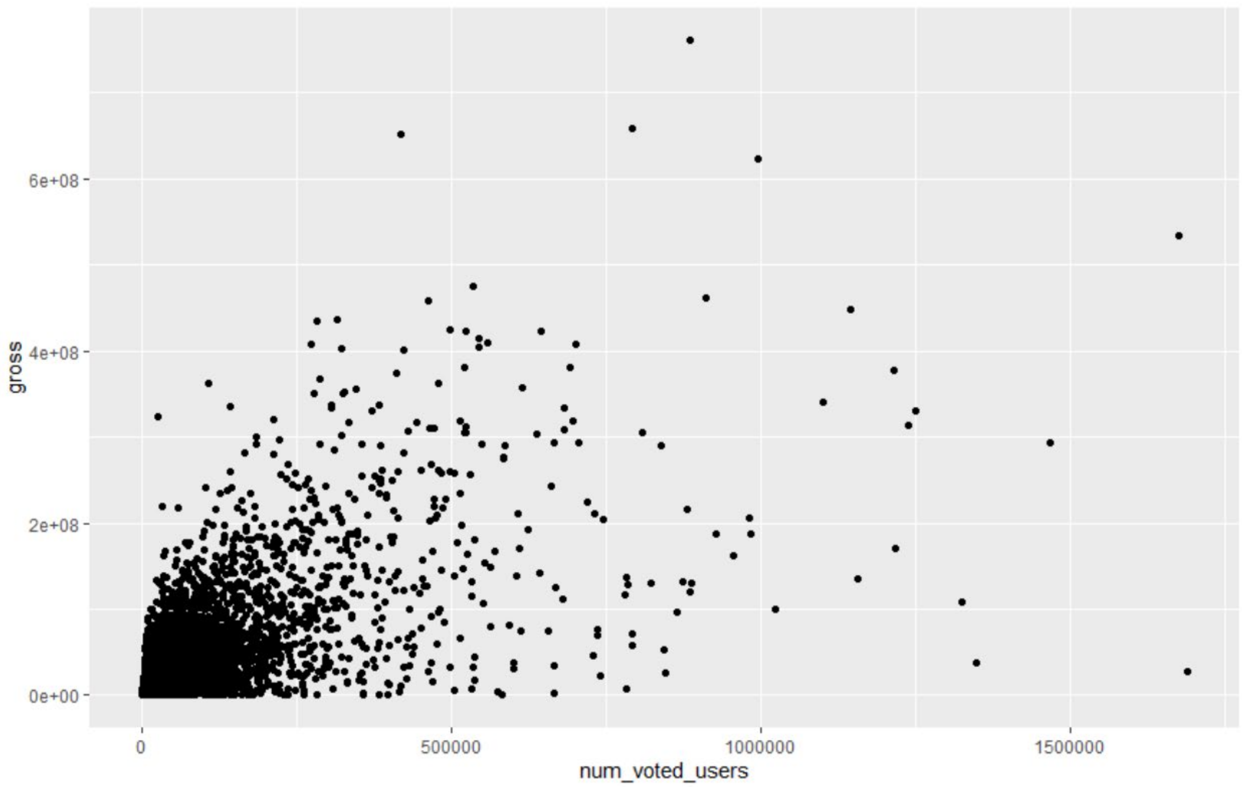
Name:_____________

Download the dataset "movie.csv" from Blackboard. Then, based on this dataset, answer following questions. You are strongly encouraged to use R.

1. How does the distribution of "gross" look like? What if you do a log transformation (create another column "log(gross)")? You may draw two histogram plots and compare them.

2. The column "num_voted_users" is the number of people who has rated that movie on IMDb (not necessarily wrote a review). Do you think it is positively correlated to the gross? Why? Test your thought quantitatively and graphically. <u>Please briefly describe/interpret your results.</u>

3. In question (2), you might have drawn a scatter plot to investigate the relationship. The plot may look like the figure below.



Did you see any problems with this figure or anything worth further and detailed analysis? If yes, what would you do? Please try your best to work it out. Briefly describe your analysis and findings. You may use following **outline** to answer this question.

From the scatter plot, we can tell… (any useful information you see at first glance).
In addition, the figure does show some interesting/abnormal patterns. For example, ……
We think it would be worth to… (propose your further analysis)
We did…… (modify data/figures)
We get …… (new figures or tables)
Now, we can tell that …… (new and additional insights you get from your analysis)

4. Looking at movie genres, can you find the most and least popular genre? Can you find the frequency of all different genres? You may use a frequency table and a bar chart to show this. Please briefly describe/interpret your results.

5. Do you think the popularity is positively correlated to the rating scores? In other words, do the movies with more popular genres have higher rating scores? Test your thought quantitatively and graphically. Please briefly describe/interpret your results.

6. Now consider another variable “budget”. Along with previously considered variables: “gross”, “imdb_score”, “num_voted_users”, and “genres”, conduct your own analysis to explore how they are correlated. Follow a similar flow as before (graphically and quantitatively), and try your best to tell different stories you uncovered from the data. Write 2-3 paragraphs as an EDA report (like a mini project). We will select the best two teams to present in class (extra credit will be given).

   Potential questions include but not limit to:

   a. Does high budget imply high gross?
   b. Does high gross imply high rating score?
   c. What types of movie requires high budget? What types need less?
   d. How does “genre” and “budget” together affect “gross”, “imdb_score” and “num_voted users”?