Assignment – SQL and R

Choose six recent popular movies. Ask at least five people that you know (friends, family, classmates, imaginary friends if necessary) to rate each of these movies that they have seen on a scale of 1 to 5. Take the results (observations) and store them in a SQL database of your choosing. Load the information from the SQL database into an R dataframe.

This is by design a very open-ended assignment. In general, there's no need here to ask "Can I…?" questions about your proposed approach. A variety of reasonable approaches are acceptable. You could for example access the SQL data directly from R, or you could create an intermediate .CSV file. I should be able to generate the SQL table(s) and data from your provided code—if you use a graphical user interface to create and populate tables, it should have a mechanism to generate corresponding SQL code.

This assignment does not need to be 100% reproducible. You can (and should) blank out your SQL password if your solution requires it; otherwise, full credit requires that your code is "reproducible," with the assumption that I have the same database server and R software.

Handling missing data is a foundational skill when working with SQL or R. To receive full credit, you should demonstrate a reasonable approach for handling missing data. After all, how likely is it that all five of your friends have seen all six movies?

You're encouraged to optionally find other ways to make your solution better. For example, consider incorporating one or more of the following suggestions into your solution:

- Use survey software to gather the information.
- Are you able to use a password without having to share the password with people who are viewing your code? There are a lot of interesting approaches that you can uncover with a little bit of research.
- While it's acceptable to create a single SQL table, can you create a normalized set of tables that corresponds to the relationship between your movie viewing friends and the movies being rated?
- Is there any benefit in standardizing ratings? How might you approach this?

You should post any code (e.g. SQL and R Markdown) in a GitHub repository, and provide a link in your assignment submission. For this assignment, you are not required to post your code to rpubs.com.

You may work in a small group on this assignment. If you work in a group, each group member should indicate who they worked with, and all group members should individually submit their week 2 assignment.

Please start early, and do work that you would want to include in a "presentations portfolio" that you might share in a job interview with a potential employer! You are encouraged to share thoughts, ask, and answer clarifying questions in this week's "R and SQL" forum.

(Optional) Reading related to this assignment

- James Le, "The 4 Recommendation Engines That Can Predict Your Movie Tastes", May 1, 2018. https://towardsdatascience.com/the-4-recommendation-engines-that-can-predict-your-movie-tastes-109dc4e10c52 This a nice backgrounder on movie recommendation engines. We'll learn more about recommender systems later in the course.
- Steve Blank, "The Customer Development Process. 2 Minutes to See Why", Jul 29, 2014. https://www.youtube.com/watch?v=xr2zFXblSRM&t=27s. In this [<3 minute] YouTube video "lean startup" founder Steve Blank talks about the importance of getting out of the building to talk to customers. I'd encourage you to adopt this "builder mentality" in your own data science work whenever it's practical, by collecting data yourself, whether it's related to a "business experiment" or a "scientific experiment."