Project Guidelines for Opportunities and Challenges of "Big Data"

A project is required for Biostat 202. The project will illustrate the steps of a data science investigation including: accessing the data, merging and cleaning the data, creating descriptions of the data with tabular and/or visualization techniques, and using a machine learning algorithm to analyze the data. Individual components of the project will be due during the quarter (weekly on Mondays by midnight starting 8/14) as well as a final description. Individual components are graded only as 0 (not turned in on time), 1 (partially completed), 2 (completed). The final description will be graded by letter grade (A+, A, A-, B+, B, B-, C, D, F).

Component 1: Specify the question, access and clean the dataset.

Due date: Midnight, Monday 8/14.

Description: Using one of the course datasets (see the CLE for the datasets and their descriptions as well as sample questions), pick a question of interest for your course project. Access the dataset you propose to use for your project.

Details: List the course dataset you will use and specify the question of interest. If your focus will be on prediction give the variable to be predicted. If your focus will be on finding structure in the data, briefly describe what you hope you will find. Describe the steps you went through to read in the data, including any problems you encountered Provide a small table describing some of the key variables (mean, minimum, maximum and standard deviation for numeric variables and frequencies of categories for categorical variables). Write up your description in two pages or less.

Component 2: Describe the data.

Due date: Midnight, Monday 8/21

Description: Using both a table as well as a figure or graphic, describe the data in a way that informs the question of interest.

Details: Provide a table and an interpretation of the results in the table. Provide a figure or graphic and an interpretation of the results (no more than two pages total).

Component 3: Analyze the data.

Due date: Midnight, Monday 8/28

Description: Apply one of the big data analytic algorithms to the chosen dataset to answer the question of interest.

Details: Describe the steps you went through to analyze the data, including any problems you encountered, why you choose that method and what steps you took to avoid overfitting (where appropriate). Maximum of one page.

Final project: Combine components 1 through 3.

Due date: Midnight, Tuesday 9/5

Description: Provide a brief write-up of all components on your project.

Details: Define the key issue, question or hypothesis you investigated. Explain the algorithm you used and why you chose it, including any problems or difficulties and how you handled them. Provide a summary of relevant results in tabular or graphical form and a discussion of what they indicate about your question of interest. Include a brief conclusion. Discuss limitations to the analysis including possible sources of bias and overfitting. **The text should be 1500 words or less and the project should have no more than 3 tables or figures.** Include a title and references but no abstract.