Assignment

Below you will find a series of questions you should answer based on analyzing a data set using R. The data file is also attached to the Toledo assignment. You should make a report of your analysis and upload this together with your code so I can run through the steps of your analysis. Please be both complete in your replies as well as concise. The assignment should be done individually this year. The university's TurnItIn software will be used to control for plagiarism in answers and code.

The dataset contains the expression levels of 71 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning. The eight classes of mice are described based on genotype (control or trisomic), behavioral assay (stimulated to learn (context-shock) or not (shock-context)) and treatment (saline or drug Memantine). The original experiment aimed to test the effect of the drug Memantine in recovering the ability to learn in trisomic mice.

1. Study and describe the data. Do you see indications of potential issues when statistically modeling the data? Explain.
2. Train and compare Ridge and LASSO models to separate the C/S from S/C (the Behavior variable) samples based on protein expression. Interpret the results of the optimizations; *at least* discuss:
    1. Do correlations between variables influence the results? How?
    2. Explain the shape of the trend in the cross-validation plot for selecting the optimal lambda value.
    3. Can a reduced set of variables predict the Behavior variable?

3. Build a boosting model and compare it to the ridge and lasso models. Are the same variables important for the predictions? Do you see evidence for non-linear effects or interactions between the most important predictor variables?
4. Does Memantine injection influence protein values when controlling for genotype and treatment?