Statistical Methods for Data Science (Spring 2018)

Mini Project 6

Instructions:

- Due date: April 26, 2018.
- Total points = 20.
- Submit a typed report.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members (if applicable)
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. Consider the crime data stored in `crime.csv`. We would like to understand how murder rate is related to the other variables in the dataset. Note that `state` is the "subject" here; it's not a predictor, and `region` is a qualitative variable.

   (a) Fit a multiple linear regression model to predict murder rate based on the other variables. Perform model diagnostics to check assumptions and perform any transformations needed to obtain a model that is reasonable with respect to the standard assumptions for linear models.

   (b) Reduce your model by removing any unimportant variables (if such variables exist). Interpret the reduced model, including coefficients and r-squared. Perform a statistical test that compares the full model to the reduced model. Clearly state the hypotheses associated with this test and interpret the results.

   (c) Use your final model to predict murder rate of a state whose predictor values are set at the average in the data for a quantitative predictor and the most frequent category for a qualitative predictor.