CON 2012
Assignment 4 (150 points)
Due: 11:59pm, April 20 (Wednesday)

Instructions: Write code in `R` to answer the following questions. Clearly label your answers using the comment function and turn in the `R` **script file** only. If there anything to discuss, please do so in your R script file.

Predicting wine preferences: the wine data (`redwine.csv`) consist of red wine samples obtained from the Portuguese "Vinho Verde" wine. The data set contains measures of wine attributes obtained at the wine certification step (measured by sensing devices) and the quality score presented by six wine experts. Our goal in this project is to build a model that predicts wine's quality using its observable attributes. The estimated model will be used to predict consumer response to a new wine and to understand which attributes need to be manipulated to induce more favorable consumer response. The data consist of the following quantities:

- `fixed.acidity`: a measurement of the total concentration of titratable acids and free hydrogen ions present in the wine (low acidity results in a flat and boring wine).
- `volatile.acidity`: a measure of steam distillable acids present in a wine.
- `citric.acid`: one of the many acids that are measured to obtained fixed acidity.
- `residual.sugar`: measurement of any natural grape sugars that are leftover after fermentation ceases; in theory, residual sugar can help wines age well.
- `chlorides`: the amount of salt in the wine.
- `free.sulfur.dioxide`: prevents microbial growth and the oxidation of wine.
- `total.sulfur.dioxide`: used as a preservative, antioxidant, and antibacterial agent in wine production.
- `density`: measure of density of wine.
- `pH`: value for pH.
- `sulphates`: a wine additive acting as antimicrobial agent and antioxidant.
- `alcohol`: the percentage of alcohol present in the wine.
- `quality`: a quality score of 3 to 8 presented by six wine experts; each expert submitted their own rating, and the median value is taken; the higher the better.

Please respond to the following questions in order:

1. Report descriptive statistics of the data and see if there is any evidence of missing or miscoded values. If any, address them properly (if none, just move on). (**10 pts**)
2. Get a correlation matrix of all variables in the data. Report correlations to two decimal places. (**5 pts**)
3. Create a set of scatter plots where the vertical axis is `quality` and the horizontal axis is each predictor. (**5 pts**)
4. Estimate a regression of `quality` on one predictor and report the results. The predictor here is the variable with the highest correlation with `quality`. (**5 pts**)
5. Estimate a regression of `quality` on three predictors and report the results. The predictors here include three variables with the highest correlation with `quality`. (**5 pts**)
6. Estimate a regression of `quality` on all other variables and report the results. (**5 pts**)
7. Conduct a 4-fold cross validation on the models you've estimated in Questions 4, 5, and 6 and report their CVE values. If all done correctly, the CVE should be the highest for the one estimated in Question 4, and the lowest for Question 6. (**25 pts**)
8. Find one alternative model fit that yields lower CVE than the ones you've estimated in Question 7. Present it's regression results and CVE (hint: consider feature selection, polynomial regression, or taking log on some of the predictors). (**20 pts**)
9. Interpret the coefficient estimate on `alcohol` (i.e., beta estimate associated with `alcohol` variable). (**10 pts**)
10. Let's say your company developed a new wine, and you want to figure out how consumers will react to it. The lab examination shows that the new wine has fixed acidity of 8.35, volatile acidity of 0.5, citric acid of 0.25, residual sugar of 2.6, chlorides of 0.08, free sulfur dioxide of 15.5, total sulfur dioxide of 45, density of 1, pH of 3, sulphates of 0.7, and alcohol of 10.6. If we assume that the model in Question 8 is your final model, what will be the predicted quality score of the new wine? If we say that the quality score of 6.5 or higher is considered "favorable" response, and the quality score less than 6.5 is "unfavorable" response, what will be its categorization? (**15 pts**)
11. How much do you trust this prediction from the estimated model? Discuss. (**15 pts**)
12. Suppose you presented this finding to the managerial staffs, and they ask you to provide suggestions on how to improve consumer response. Based on your analyses so far, what would you suggest? (**15 pts**)
13. Think about the data carefully. I feel that there is one critical problem that makes the actual consumer response very different from what we see in the analyses. What is the problem? Discuss. (**15 pts**)