Regression Project
Due Monday May 13th

Description of Dataset

The following synthetic dataset describes the performance of lawyers at a particular law firm last month. The lawyers are judged by the number of cases that they handle in a given month. The variables we have are

- CTM: Cases this month
- CLM: Cases last month
- AGE: The lawyers age
- SDY: The number of sick days that the lawyer took in the last year.
- LVL: The lawyers' rank within the firm. From most junior to most senior the ranks are:
  - Associate
  - Senior Associate
  - Junior Partner
  - Senior Partner
  - Named partner

Our goal in this assignment will be to predict the number of cases a lawyer will handle this month given the number that they handled last month, their age, their level of seniority and the number of sick days taken in the last year.

1. **Single Variable Regression:**
   Use single variable regression to predict CTM from each of

   (a) CLM
   (b) AGE
   (c) LVL
   (d) SDY

   For each model comment on how well it fits the training data and the test data. What is the training R-squared. Draw scatter plots with a line to show the fit. [20]

2. Several Variable regression/ Use the techniques of multiple variable regression to fix a model using some of the variables given. Which models work best? Demonstrate iterations involving different choices of variables. Explain which ones worked best and how you selected them. Show plots of errors. [20]

3. For each model fitted discuss weather it improves on the previous one and why justify the choice of the subsequent model. [10]

4. Discuss if the models chosen make sense [15]

5. Use whatever heuristics you feel most appropriate to support the models (scatter plots, AIC/BIC and so on) [15]

6. Try to engineer features that generate better performing model overall. Explain why you chose these features. [10]

7. Discuss if the models chosen make sense [10]