Stat 123 Homework Assignment 4
Due Friday April 8th by 9:00pm

Using R Markdown, please complete the following assignment. Your assignment should be submitted as a pdf (whether you knit directly to PDF, or knit to HTML or Word and then convert the file to a pdf).

1. Download and save the *AdmissionsPredict.csv* dataset and read it into R. This data set consists of data regarding international students who are applying for graduate programs in English speaking countries. The data set contains 7 variables:

   - GRE (Graduate Record Examination score)
   - TOEFL (Test of English as a Foreign Language score)
   - University Rating (score out of 5)
   - SOP (Statement of Purpose, score out of 5)
   - LOR (Letter of Recommendation, score out of 5)
   - UGPA (Undegraduate Grade Point Average, score out of 10)
   - Chance of Admit (value between 0 and 1)

   (a) The response variable is y = Chance of Admit. All other variables are possible explanatory variables. Create a vector called *xnames* which contains the names of each of the explanatory variables.

   (b) Use the command $par(mfrow = c(2,3))$ and then write a for-loop which plots each of the explanatory variables against the response variable. Make sure each plot has an x-axis title and a y-axis title. Use *xnames* from part (a) to create the x-axis title.

   Hint: You did a very similar question in lab with Steve.

   (c) For which explanatory variables are you able to identify the form of the relationship with y? What is the form that you see?

   (d) Create a linear regression model called $full\_model$ which includes all of the possible explanatory variables. Write out the model that you obtain.

   Example: $y = 0.3 + 0.1(x1) - 0.5(x2) + ...$

   (e) Are all terms significant? Identify any variables which should be removed from the model (and show the code that you are using to make this decision).

(f) Create a new model called *new_model* which contains only the terms which were significant in the full model. Write out the model that you obtain.

(g) What is the range of values for each variable included in the *new_model* that we can use for prediction (so that we avoid extrapolation).

(h) Consider a student with a GRE score of 320, a TOEFL score of 101, applying to a University with a rating of 4, with a SOP score of 3, a LOR score of 4 and an undergraduate GPA of 8.4. Use your model from part (f) to predict this students chance of being accepted to the graduate program at their University of choice.

2. Type in the following vectors which represent people of various ages (in years) who are each timed (in seconds) running the same distance.

```
age = c(2,3,4,5,8,11,14,17,21,28,38,50,67,83)

speed = c(65,58,40,37,32,26,18,16,17,17,23,29,42,59)
```

(a) Which is the response variable and which is the explanatory variable?

(b) Plot the variables (you do not need any titles). What form does the relationship seem to have?

(c) Fit a model to the form that you identified in part (b). Write out the model that you obtain.

(d) Use your model to predict how long it would take for a 70 year old to run that distance.

(e) What percentage of the variation in the response variable can be explained by the variation in the explanatory variables in the model?