EDA Homework 4

Due: Friday, October 12 at 5pm.

Work in groups of up to 3, and make one submission for each group.

The file movie_budgets.txt contains data on the budgets of 5,183 movies from 1906 to 2005, along with their lengths in minutes. Read your data into R as a data frame called movie_budgets. We wish to study log10(budget) as the response variable and year and length as explanatories. Note that these movies are not a representative sample of all movies, so we're not trying to generalize, only describe the data we have.

- Using loess or otherwise, fit a model to predict log10(budget) from year and length. For simplicity, do not transform year and length (even though a transformation of length would probably be sensible.) You will have to make a number of modeling choices:
  - Should you fit a linear or curved function for year?
  - Should you fit a linear or curved function for length?
  - Do you need an interaction between year and length?
  - What span should you use in your loess smoother?
  - Should you fit using least squares or a robust fit?

  Some of these choices are clear-cut, while others will be a matter of preference. Either way, you must justify all your choices.
- Draw one set of faceted plots to display the fit – either condition on year or length, whichever seems to you to be more interesting. Choose a sensible number of panels. Briefly describe what this set of plots shows you.
- Draw a raster-and-contour plot (or other "3D" plot of your choice) to further display your fit. Briefly describe what, if anything, this plot shows you that your plot for question 2 didn't.

What to submit:

- An R code file to reproduce your fit and plots.
- A write-up containing:
  - The line of R code you used to produce your loess model.
  - A justification of your modeling choices.
  - Two graphs and brief comments on those two graphs. (You should draw many more than two graphs when deciding what model to fit, but only include two graphs in your submission.)