Your task in this project is to select a real-world data set to which multiple linear regression (or a related method) can be appropriately applied, and then to conduct a statistical analysis (including diagnostics of course) of the data set using the techniques that you have learned this semester, as well as hopefully some additional method(s) that you learn or apply for the first time in this project. The analysis should be your own (with computations conducted in R), even if you select a data set from a book or paper with its own analysis. Do not do the same analysis that is posted on a website. It is okay if your source has an analysis as well, but your analysis should be different from it (e.g., you could use a different response variable or different data transformations). I would recommend that you choose a data set that you find interesting or relevant to an area of study that you enjoy. Hint: Try consulting faculty in your department to see if they have any suitable data sets!

As before, the idea in this project is to pretend that you are a statistical consultant. It is up to you both to conduct a thorough analysis of that problem (or as thorough as possible given the material that we have studied) and to communicate your results to your clients (who I will pretend to be). This communication is an important part of statistics. You may assume some familiarity with statistics on the part of your clients, but you should not get completely mired in technical terminology.

Projects will be done individually and will consist solely of a paper (no presentation for this project).

The schedule of events for these is as follows:

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, December 11</td>
<td>Topic proposal due (having run some diagnostics)</td>
</tr>
<tr>
<td>Friday, December 21</td>
<td>Paper due</td>
</tr>
</tbody>
</table>

Since the semester ends on December 21, papers will not be accepted after that date. You should put a hard copy of the paper under my office door and email me to let me know that it is there. Please do not email me your paper; I will not accept in this form. If you are leaving early, you need either to turn it in before you leave or find a way to get a hard copy to me by December 21.

**Topic proposal:** As with earlier projects, in about a paragraph, describe the data set that you are proposing to analyze: where it comes from, how it was collected, why you selected it, and so forth. In addition to what you did for the topic proposal for Project 1 though, you should have run the relevant diagnostics to make sure that multiple regression analysis is suitable for the data. (The regression assumptions should be satisfied, etc.) You do not have to include graphs in the topic proposal, but you should state what diagnostics you have done and why you therefore know that regression analysis is appropriate. I won’t collect these but will simply ask you about them in class.
**About the paper**
The paper should include the following sections: introduction, data set background, data, *statistical background*, diagnostics, analysis, conclusions, and bibliography. All of these should be clearly labeled, except for the introduction, which simply begins the paper without a header.

In the statistical background section, you should explain a bit about any new methodology or statistical techniques that you learned for this paper, or that you are applying for the first time in this paper. If you are using only things covered in class, you should at least write out a section explaining the most recently learned or most advanced part of the statistics that you are using.

Although it is not usual for a paper in a journal, in this paper you should include the relevant graphs and statements about how you know this data set is suitable for multiple linear regression. This should be modeled after the diagnostics that you have done in the homeworks (as in the text and the Faraway handout).

As with Project 3, as part of this paper, I would like you to submit an R script file that can be run with the `source()` command. This script file should contain comments indicating what each command is designed to do. We have discussed how to write such a file in class, and I would be happy to help anyone who is still having difficulties writing the script. You should email me this script, and you should also print out a hard copy and turn that in with your paper.

As with previous papers, it is a good idea to have someone who has not taken the course read your paper and see which parts they understand and which they don’t. If they don’t understand even the general idea of what you have shown, then your paper is not clearly enough written. Likewise, if they don’t understand the techniques you are using, they should still pretty well be able to follow along and understand what you are doing and why.

**A brief comment**
Please see me at any point along the way if you would like any help or guidance in selecting a topic or conducting your analysis. While I certainly won’t do either one for you, I may be able to get you unstuck if you happen to get stuck at some point (including finding a topic).

**A word of advice**
As you write your paper, keep the following principle in mind: *The data tell a story (or even many stories). You are to find that story and put it into words and pictures so that it can be understood be the reader.* When the reader has finished your paper, he or she should be very clear on what questions the paper addressed, what answers it proposes, and what the difficulties and drawbacks of these answers are.

*Thanks for an enjoyable semester!*